An Analysis of Qualitative Feel as the Introspectible Subjective Aspect of a Space of Reasons

Michael James Stuart Beaton

Submitted for examination in the degree of Doctor of Philosophy

at the University of Sussex, April, 2009

Declaration

I hereby declare that this thesis has not been and will not be submitted, in whole or in part, to this or any other University for the award of any other degree.

Signature: .....

An Analysis of Qualitative Feel as the Introspectible Subjective Aspect of a Space of Reasons

Michael James Stuart Beaton

Summary

This thesis presents an analysis of qualitative feel ('qualia'), based on a Sellarsian 'space of reasons' account of the mental. The first non-introductory chapter, Chapter 2, argues against an over-strong phenomenal realism (the claim that inverted spectra, zombies, etc., are at least conceptually possible), and against the modern phenomenal concept defence of such claims. Nevertheless, it is agreed with the proponents of these views that we must allow for introspective knowledge of our qualia, if we are to take qualia seriously at all. It is therefore proposed that we allow our search for qualia to be guided by some independently plausible theory of introspection. In Chapter 3, Shoemaker's account of introspection is presented, extended in certain respects, and defended against some current objections. Chapter 4 is used to argue that Shoemaker's current account of qualia can only be made compatible with his account of introspection by paying certain very high costs (which Shoemaker is aware of, but seems willing to pay). However, it is also argued that Shoemaker's account of qualia has some attractive features, which can be preserved. In Chapter 5 a novel analysis of qualia is presented, as non-intrinsic (i.e. relational), introspectible aspects of mind, fully capturable at the level of a 'space of reasons' analysis of an agent. A detailed analysis is given, for the cases of colour qualia and of pains. The aforementioned, attractive features of Shoemaker's account are adapted, in order to address some of the complexities of the different ways in which we can think about such qualitative properties. In Chapter 6, it is argued that this account of qualia has the potential to explain plausibly many of our problematic intuitions concerning qualia including: their ineffability; our ability to know them infallibly and incorrigibly; and (though only in weak senses) their intrinsicness and privacy.

Submitted for examination in the degree of Doctor of Philosophy
at the University of Sussex, April, 2009

Acknowledgements

I would like to thank my two supervisors, Steve Torrance and Ron Chrisley, for their encouragement and support. I would also like to thank Steve for deciding to take a chance on me as I returned to academia after an extended break.

Amongst my fellow graduate students over the years at Sussex, I would like to thank the following for many discussions and for invaluable feedback on written work: Tom Beament, Rob Clowes, Chris Davia, Chrisantha Fernando, Tom Froese, Hanne De Jaegher, Miriam Kyselo, Rowan Lovett, Marek McGann, Simon McGregor, Tony Morse, Will Newhouse, Joel Parthemore, Alexandra Penn, Marieke Rohde, Nathaniel Virgo and Alexandros Zographakis. Thanks, too, to all the other members of the various research, reading and discussion groups at Sussex who have given feedback and generally created a friendly and intellectually stimulating research environment, including all the members of PAICS, E-Intentionality, CogPhi, CCNR, Life and Mind and PAC Lab.

I owe Inman Harvey here at Sussex more than, I think, he realizes. I am deeply indebted to him for some two or three years of patient conversation, during which he successfully convinced me that there was something fundamentally wrong with the representationalist assumptions underpinning most theorizing in cognitive science; and – though I hesitate to mention this – he also made sure that I realized that I needed to be much more careful in my thinking, as I moved between mind talk and brain talk. This would not have been the thesis it is, without his input.

I would like to thank Lucy Allais for running the graduate philosophy class which introduced me to McDowell.

Many established academics from other Universities have been extremely generous with their time, at one point or another, in allowing me to work through details of their ideas (or mine!) by email, or during face to face conversation, including: Torin Alter, Uzi Awret, Ned Block, Patricia Churchland, Daniel Dennett, Marco Giunti, Germund Hesslow, Nicholas Humphrey, Dan Lloyd, Thomas Metzinger, Alva Noë, Chris Nunn, Kevin O'Regan and David Rosenthal.

Acknowledgements

My thanks to David Chalmers for deciding that there was something worth saying in the first version of my RoboDennett paper (but also for helping me to understand how much needed doing to turn it into a publishable paper).

I would especially like to thank Romi Nijhawan and Beena Khurana for providing an extremely constructive and supportive atmosphere, within their research group, into which I was welcomed as I explored psychophysics for a period of around a year, during the course of my DPhil research. (Romi, we *will* get that red-green flash-lag data written up as a journal paper, now that I have submitted!) My thanks to Zoltan Dienes for also providing help and guidance in this area.

Very grateful thanks to my sister, Caroline, for several instances of specific and invaluable support, as I tried to work out how to manage my time and move forward with this research; and, for that matter, for helpful tips on public speaking!

This work was financially supported (during the second and third of the three full-time years allocated for DPhil research) by a Graduate Teaching Assistantship awarded by the former School of Cognitive and Computing Sciences (COGS) at the University of Sussex. COGS (formerly) and Informatics (latterly) have also been generous in provision of funding to attend various academic conferences over the years. The James S. McDonnell Foundation and the European Science Foundation were kind enough to provide funding to attend two separate but equally invaluable Summer Schools, each of which brought together early-career and established researchers from philosophy and neuroscience.

Preface

The bulk of Chapter 2 (specifically, Section 2.2) is in press elsewhere (Beaton, in press). A large portion of Chapter 6 (Section 6.4) is already in print (Beaton, 2005). These papers were produced as an integral part of the present thesis research. Copyright in these papers has been assigned to the publishers of the *Journal of Consciousness Studies*, and they are included herein with permission for non-commercial distribution only.

Table of Contents

1. Introduction

1.1 Introductory Remarks

This thesis is centrally concerned with phenomenal qualities, or *qualia*: the properties of a conscious experience which determine what it is like (Nagel, 1974) to have that experience.

There is a standard conception of such properties on which I can know them in introspection. For instance, I can introspect the phenomenal redness of my reds (how red objects look to me) and I can wonder whether or not I am thereby introspecting the same property which you introspect, when you introspect the phenomenal redness of your reds. This thesis aims to support and naturalise that conception; but it aims to do so in a very nuanced way, navigating the treacherous passage between the Scylla of eliminativism and the Charybdis of dualism.

Dennett (1988; 1991) has argued that there are no such introspectible properties. He correctly observes that my red might well pick out slightly different aspects of the world from your red. Armchair philosophy might indicate to us that such differences would be likely to arise between subjects, purely on the basis of 'nurture' (upbringing, differential experience). But we can do better than that. For such differences are empirically likely to be present, between any typical pair of human subjects, on the basis of 'nature' alone: there is clear evidence of small, genetically based variations in peak colour-cone spectral sensitivity amongst humans with 'normal' colour vision (Jacobs, 1996)[1]. That's all well and good. But, Dennett alleges, "that idiosyncracy is the extent of our privacy" (Dennett, 1988). Dennett is claiming that when you've described this kind of difference between subjects, you have said everything which there is to say in respect of how they differ in their subjective perceptual response to red.

This thesis aims to naturalise something both more private and more subjective than that. I will argue that the intuition that there are qualia is inseparable from the intuition that the introspectible, phenomenal properties of experience might vary, even between two subjects who are seeing *exactly* the same property of the world (a colour, say) *as*

[1] Two subjects who differ in this way (and who are using their colour cones optimally) will not be able to make exactly the same colour discriminations as each other.

exactly the same property of the world, and who agree (in a shared, public language) *that* they are seeing the same colour as the same colour.

To wish to naturalise something which is introspectible, and which might vary even whilst all that is fixed, is to ask for a lot; but it is what I aim to do. Nevertheless, it is quite possible to ask for still more; to ask for too much. Many (and historically, most) "qualia freaks" (Jackson, 1982) have supposed that intuitions such as the above can only be naturalised by showing that qualia are, in principle, separable from behaviour. That is, it is supposed that we could, at best, have *a posteriori* (i.e. empirical) reasons for associating given qualia (which is the plural, the Latin for 'qualities'; *quale* is the singular) with given behaviours, in a given population. To make the same point in the converse direction, it is assumed that to correctly allow for qualia at all, we have to allow that there is nothing about their nature such that there is any *a priori* (analytic, logical) link between a given phenomenal feel and any particular kind of behaviour.

What was surprising, to me, as I researched this thesis, was to realise that the above view (that there are qualia, but that they can only be related *a posteriori* to public facts) has been very common amongst physicalists over the years (e.g. Shoemaker, 1975; Lewis, 1980; Churchland and Churchland, 1982). Recently, this trend has become fully explicit within analytic philosophy, with many philosophers (e.g. Loar, 1997; Papineau, 2002; Carruthers and Veillet, 2007) attempting to analyse this allegedly *a posteriori* relation between qualia and public behaviour, and to show that nothing therein need threaten physicalism.

All of this, I will argue in some detail, is deeply mistaken. If the relation between phenomenal qualities and public facts is *a posteriori*, as so many self-styled physicalists have alleged, then qualia cannot be situated in the physical world using anything like the normal scientific mode of explanation. At best, this would allow us a very 'thin' ontological form of physicalism, with none of the explanatory benefits that physicalism is supposed to bring. At worst (and as I argue) Chalmers (1996) has been right all along: *if* qualia essentially have this *a posteriori* relation to the public world, then qualia are no part of the world as physics understands it.

Of course, Chalmers (1996; 2006) concludes from the above that, since there are indeed qualia, qualia are no part of the world as objective physics understands it. And it can be very easy to feel that the only options available are this conclusion or some broadly Dennettian eliminativism about qualia. Here, I try to navigate the difficult waters already mentioned between these two extremes, both of which I believe are

unacceptable: I develop and present an analysis of qualia, even whilst denying what the vast majority of qualia freaks have claimed – that qualia are logically separable from behaviour.

1.2 Chapter Overview

1.2.1 Chapter 2 – Background Issues

In Chapter 2, I give the above mentioned arguments to the effect that normal scientific explanation is entirely incompatible with there being an intrinsic, mental aspect to qualia which can only be related *a posteriori* to public mental facts[2]. In addition, I observe that there is copious disagreement as regards which properties, or even which kinds of properties, are accessible in introspection. Furthermore, this disagreement is at its worst specifically as regards the introspectible accompaniments of, or properties of, perceptual experience. But, of course, qualia are (amongst) the introspectible accompaniments of, or properties of, perceptual experience. I therefore suggest that the premise that qualia have this *a posteriori* relation to public facts amounts to an (at least implicit) endorsement of a *theoretical* commitment about introspection, rather than being something which we can pre-theoretically know, about the nature of our phenomenal qualities.

I suggest that we should instead allow ourselves to be guided, in our search for qualia, by whatever our best independently plausible theory of introspection is. I therefore propose a minimal definition of qualia, as those introspectible properties which can vary as described (i.e. even as between two agents who are seeing the same public property as the same public property, and can agree that they are). I accept that, were there to be no properties matching this definition, I would be forced to agree that there are no qualia.

To proceed, we will need an independently plausible account of introspection. Much of the rest of the thesis, including the account of introspection to be offered in Chapter 3, is predicated on the notion of mind as physical locus of action for reasons (Sellars, 1956; McDowell, 1994; Hurley, 2003). Therefore, before I can present the account of

[2] In fact, I have to allow that a *highly* reductive form of 'explanation' is still possible. I suggest that this would be an extremely undesirable outcome, and also argue that the degree of reduction required to explain qualia, in that case, would be greater than that involved in explanation of normal macroscopic properties such as liquidity and heat.

introspection defended in Chapter 3, there is a final piece of stage-setting work to be done in Chapter 2, with a brief introduction to, and clarification of, this notion of mind as locus of (at least counterfactual) rational action.

With this done, the stage is set: if mind can be analysed thus; if an account of introspection can be found using this analysis of mind; and if properties satisfying the above definition of qualia can be found within the properties introspectible on such an account, then we would have plausible candidates for qualia which are (in virtue of being thus situated) not logically separable from publicly accessible facts, because not logically separable from their role in (at least counterfactual) action.

1.2.2 Chapter 3 – Introspection

In Chapter 3, I present and defend an analysis of introspection introduced by Sellars (1956) and defended in detail over the years by Shoemaker (1996). According to this analysis, introspection should be understood as a single-step, non-inferential rational transition, where the concepts employed in introspection are the very same concepts employed in public mental ascription. An example of such a transition would be the transition from seeing a red ball to the state of thinking that (or being aware that) one is seeing a red ball (with 'seeing' here understood on a public, at least counterfactual behavioural basis).

Shoemaker argues in detail (in the case of many specific examples) that we can't be rational and wrong in self-ascription of such public mental concepts. This, he further argues, throws into doubt the claim that we need anything other than rationality for introspection. In particular, Shoemaker is keen to call into question any perceptual or quasi-perceptual model of introspection[3]. I present and defend Shoemaker's arguments for this.

I try to head off an easy misreading of Shoemaker, for Shoemaker presents detailed lines of reasoning designed to show why such transitions are rational (in many different, specific, cases). It is easy to suppose that Shoemaker requires of an agent that it be able to understand such lines of thought (at least in some informal way) in order to introspect. But Shoemaker does not require this; he only requires that an agent *make*

[3] In this thesis, I will follow the widely adopted convention of continuing to call this process introspection, even in the case where one holds that this is etymologically misleading.

such transitions, when and because they are rational. This is a subtle point which I explore in some detail.

If one claims that introspection simply involves 'rationality', it can look as if one is entirely ignoring empirical questions about what is required in the physical constitution of an agent, in order for it to introspect. I argue that this is not so. Staying at the theoretical level, I claim that what I am doing is simply getting clear on what introspection *is*, as an essential (but separable) part of the process of trying to understand how it is 'implemented' in real agents.

But I also say some things which are a little more specific than that. I note that there is a *prima facie* disagreement between Shoemaker and Sellars, exactly as regards this issue: they each talk, in similar terms, about the mechanisms which might be involved in noninferential access to internal physical states (such as blood pressure, or whether one is infected by a particular virus), but they seem to say exactly opposite things about such cases. I argue that the disagreement is only apparent. Furthermore, I argue that showing how to resolve the apparent disagreement sheds more light on what is involved, subpersonally, in the case where an agent can introspect.

Finally, I address some more recent objections to this model of introspection. Gertler (2003/2008) has suggested that it may well be "overly demanding", due to requiring too much rationality to be plausible as an analysis of introspection in the most basic case. Kind, in a related vein, argues that even if Shoemaker's arguments are correct, and we can gain self-knowledge in the way he describes, that this is still not introspection: Kind alleges that Shoemaker has mistaken an essentially third-person way of gaining self-knowledge for essentially first-person self-acquaintance.

I suggest that to respond, particularly, to Kind's objection, we have to modify Shoemaker's arguments (presentationally rather then substantively). I further argue that we can *strengthen* Shoemaker's claims, in response to Kind. In particular, I argue a) that Shoemaker has provided a genuine analysis of introspection rather than merely arguments *against* the quasi-perceptual model (which is all that he explicitly claims to have done), b) that if we take the quasi-perceptual model seriously enough to compare it to the rationality model of introspection, we find strong reasons for saying that the latter is introspection and that the former (even though, in a sense, possible) is not, and finally c) that Shoemaker's arguments concerning to nature of introspection can be generalised, to show that any aspect of a space of reasons as such is the right kind of state to be introspected in this way.

1.2.3 Chapter 4 – Shoemaker's New Account of Qualia

If the above is all and only what there is to say about introspection, then qualia must be public mental properties, if there are to be qualia at all. For qualia can (at least in some cases, at least in us who seek to explain them) be introspected, and, on the account of introspection I have just defended, only public mental properties can be introspected[4]. At this point, I ought to be ready to present my account of qualia, which operates within these constraints.

It turns out, though, that I have a rather striking problem to address first. For I am building on Shoemaker's account of introspection, and yet Shoemaker's own latest account of qualia *does not* accord with the above constraints; it still involves private, intrinsic, non-relational properties which help to determine 'what it is like' to have an experience. Surely Shoemaker can't be wrong about the implications of his own account of introspection? It looks as if I must have misunderstood (at least the implications of) Shoemaker's account of introspection, or his account of qualia, or both.

In this chapter, I present Shoemaker's most recent account of qualia, show why I object to it, and show how to resolve the above worries. Shoemaker has always believed that qualia can vary, as between subjects who are seeing the same parts of the world *as* the same parts of the world. He has also always (I believe incorrectly) assumed that the only way to properly naturalise this intuition is to allow that qualia are not fully determined by behaviour and counterfactual behaviour (e.g. Shoemaker, 1975). More recently (Shoemaker, 1994d), he has recognised that this position is in tension with the account of introspection which he has been developing over the years. But he has not abandoned his belief that qualia are only (at best) *a posteriori* relatable to behaviour. Nevertheless, he does now accept that he cannot allow that such qualia are introspectible.

Shoemaker finds a clever solution to this apparent incompatibility. He argues that we do not, in the first instance, know qualia, rather we know what he calls the 'phenomenal properties' (roughly, secondary properties) of objects. His proposal is that we see colours (say) in and by seeing relational properties of objects, such as the property of

[4] Not public in the sense that I can know what all your mental properties are, just by looking; but public in the sense that I could, in principle, find out what any given mental property of yours is, just by looking, asking the right questions, etc., without there ever being a need for me to, e.g., extrapolate to your case, starting from properties which I can only really know in a first-person way, in the first instance.

tending to cause this or that colour quale in me. On this proposal, when I see blue (for instance) I also – at the same time, combined into a single experience – see it as 'that which causes this or that qualitative property in me'.

Since I am effectively claiming that Shoemaker's new account of qualia amounts to be a rearguard defence of a flawed analysis of qualia, it might seem implausible (it did, to me) that I would find much in it to attract me. In fact, there are some aspects of it which I do find attractive and which I therefore incorporate as positive features of the alternative account which I offer, in the next chapter.

As such, I have more than one reason for presenting Shoemaker's account of qualia in a reasonable level of detail, in this chapter, and I do so.

But I argue that there are significant problems with it. We can begin to get a sense of why this might be so, by noting that the account breaches a broadly Evansian (Evans, 1982) analysis on which, in order to perceive something a certain way, one should know (or understand) what it is for something to be that way. Indeed, Shoemaker himself comes close to endorsing this formulation in his own arguments[5].

This problem is closely related to what I think is the deepest problem with Shoemaker's new account of qualia: it is incompatible with a causal account of our knowledge of these relational 'phenomenal properties', whether at the subpersonal or the personal level. Very surprisingly, Shoemaker acknowledges this (in passing, in a footnote: Shoemaker, 1994d n.7) at least as regards the subpersonal level. He would appear to think that this is a price worth paying, for the prize of naturalising qualia[6]. But it is a very high price. Do we really want to rule out subpersonal causal explanation?

However, I argue that the best way to show why this is a price which should not be considered worth paying, even from Shoemaker's own point of view, is to look at the problems which come with ruling out a causal account at the personal level. For Shoemaker is trying to defend functionalism, but in the end the things he says concerning qualia completely undermine this position. On a functional account, mental states are supposed to be analysed in terms of their causal relationship to one another, but here we have a kind of knowledge which *cannot* be analysed causally; if this is correct, functionalism would not be the correct account of the mental. Once again,

[5] As I note, the priority may be reversed, since Evans is deeply influenced by earlier work by Shoemaker.

[6] Indeed, one could easily argue (c.f. Chalmers, 2006) that modern phenomenal concept strategists are making exactly the same move.

Shoemaker is perhaps aware of this cost, but it is a very high cost. Do we really want, or need, to pay it?

Shoemaker has not – as far as I am aware – responded to the second of the two charges above. Nevertheless, having seen that Shoemaker *is* aware of at least some of the high costs in making his account of qualia compatible with his account of introspection – and having seen very good reason not to want to pay those costs – is hopefully enough to reassure worried readers that I have not misrepresented any aspect of Shoemaker's position.

That said, we can move on to the analysis of qualia which this thesis offers, which avoids the costs of Shoemaker's account by analysing qualia as introspectible, but in principle public, aspects of a space of reasons as such.

1.2.4 Chapter 5 – A Space of Reasons Analysis of Qualia

A key observation grounding the analysis of qualia given in Chapter 5 is that a space of reasons level description of the actions of an agent cannot be complete without affect. As philosophers going back at least to Hume (Hume, 1739-1740/2000; e.g. as quoted in Froese, 2009) have observed, no mere collection of facts is a reason to *do* anything. One must simply be moved, in the face of at least some situations, to do something; this cannot be reduced to, or replaced by, the appreciation of yet more facts.

For this reason, I argue that it is a mistake to think of the most basic desire-like state as propositional. To give an example: I see food, in the most basic case, by seeing it *as* a strawberry or a banana[7], say; but if I am hungry, then I also desire the food (the strawberry; the banana). However, this latter state[8], of desiring the food, is not to be analysed as a state wherein I think *that* the food is desirable. It just involves me *desiring* the food. I suggest that, when I am hungry, my space of reasons becomes modified in such a way that the *food itself* becomes a reason (for me, in that state) for certain basic

[7] This requires at least practical understanding of what it is (c.f. Section 4.3.1, Sections 5.3.5-5.3.6 and Chapter 6, footnote 153) for something to be a strawberry, or a banana: the kind of understanding required for competent, flexible, rational interaction with these things, in the context of the creature's interests (c.f. Hurley, 2003).

[8] A couple of points: firstly, this is necessarily only a *partial* state of an agent, for of course desiring something could not be a complete description of a mind; secondly (and as I also note in Chapter 2, footnote 19), I use 'state' throughout this work in a sense (which is ubiquitous in the physical sciences) which does not in any way exclude the possibility of a fundamentally process-based analysis of the 'state' in question.

actions, such as taking and eating. To put the claim even more bluntly: hunger *is* such a modification of a space of reasons.

Hunger, of course, is also a state with introspectible qualitative feel. So, can the above behaviourally based analysis be *all* there is to say about hunger? Can the above *be* the introspectible, qualitative feel, as well as a mere description of the actions taken, when I have that feel? With some caveats and clarifications (but not with any which substantively affect the conclusion), I argue that this can be, and indeed is, how things are.

One of the first steps is to note that such candidate qualia are, indeed, introspectible. For they are properties of a space of reasons as such (i.e. this property, *qua* this property, is fully defined by its role in a space of reasons), and I have already argued that any property of a space of reasons as such is the right *kind* of thing to be introspected[9].

In order to put more flesh on the bones of this analysis, I proceed to look in some detail at the most standard examples of qualitative feel: the qualia associated with the perception of public colour properties, and the quale of pain.

In the case of colour qualia, I argue that we can avoid the drawback of Shoemaker's account, wherein an agent could see something a certain way, without knowing what it is for something to be that way. On the account I offer here, in the most basic case an agent simply sees blue *as* the fully public (if gerrymandered[10]) property blue. In such a case, the agent *has* the quale associated with blue (they are affected 'bluely'), but they need not know that they have that quale, nor think of blue *as* having that effect. Next, we come to the case of a more theoretically informed agent with the (at least 'folk') concept of 'the effect which blue has on me'. Such an agent can (or rather, at least in principle could) introspect the effect in question, for it is the right kind of property to be introspected. Moreover (and here I incorporate the attractive aspects of Shoemaker's account mentioned above) such an agent could also noninferentially[11] *see* blue as having

---

[9] Nothing here says that every agent with such a property can in fact introspect it, just that some agents with such properties could do so, compatible with the most plausible account of introspection (c.f. Chapter 2, footnote 48).

[10] That is, a property having an outline which may well depend on the interests and constitution of the type of creature doing the seeing (this is Dennett's usage, e.g. Dennett, 1991).

[11] Neither perception nor introspection involve personal level inference, in the most basic cases.

the property of causing that subjective effect; though this case of 'seeing' is not purely perception and not purely introspection, it fundamentally involves both.

Although this latter possibility, of seeing colours in this way, has been incorporated into my account from Shoemaker's current account of qualia, my account of what qualia *are* differs fundamentally from Shoemaker's: in Shoemaker's account, even the least theoretically informed of us somehow, inexplicably (as I have argued), sees colours *as* having such properties; in the present account, perceived public colours *have* such properties always (they have the effects they have, on us), but an agent has to know what it is for something to have such a property (at least in some practical sense), as a necessary precondition for seeing it *as* having it.

I briefly question whether there is any good reason (other than historically misleading precedent) to call perception on such an account 'representation': for neither the theorist nor the subject need have any aspect of such a state in view, *as* a representation (in the mundane sense, in which road signs represent), in order for the nature of the state to have been fully grasped.

Next I address the issue of *pain* (the feeling) and of *pains* (which are, on the analysis to be rejected: intrinsically awful objects of direct, internal, mental awareness). The qualitative feel of pain is, on this account, the introspectible modification of a space of reasons associated with (and corresponding to motivation to do something about) at least seeming damage to an at least seeming body part[12].

However, in a move which may well not be strategically advisable, but which I believe is worth making, in order to show how this account can correctly analyse pre-theoretic intuitions, I argue that there can still be pains. Of course, I have absolutely no wish to reinstate private mental objects, and no intention of doing so. Nevertheless, from a pre-theoretic point of view, it reduces the plausibility of an account of pain to say that there are no pains, in *any* sense. Equally, it seems to me that there is no reason to say so, if pains are properly analysed.

Pains, I suggest, are (at least seeming) body parts, sensed painfully. (*Pain*, rather than *the pain* or *a pain*, is still the associated introspectible feeling.) Thus, in much the same way that the food itself is a reason for action in the case of hunger, so the body part itself becomes a reason for action (the kind of action which would typically reduce or

[12] One can, of course, act (or at least be motivated to act) *as if* one has damage to a body part, when there is no damage and even when there is no body part (Ramachandran and Blakeslee, 1998).

mitigate bodily damage). A body part in such a case is, I suggest, *a pain* (this is one of what are no doubt many meanings mixed up in the English word). On this analysis, pains cannot exist unperceived, but this need not be metaphysically worrying. For ⌜a body part sensed painfully⌝ cannot exist unsensed; but, of course, ⌜a body part⌝ sensed painfully can still perfectly well exist unsensed.

I suggest that it need not worry me unduly if there are aspects of the English word 'pain' under which some bodily damage is sometimes correctly called a pain, even when unsensed, or when sensed but not sensed painfully[13]: for there is certainly only unsensed pain in such a case to the extent that the body part in question would be felt painfully, if only certain counterfactuals obtained.

Churchland and Churchland (1982) use the observation that pain can feel different, in different cases (sharp pains, searing pains, dull aches, throbbing pains, and so on) to support the assertion that the feel of pain does indeed outrun the correct 'functional' (i.e. behavioural) characterisation of it: pains are all pains because they *share* a behavioural profile, they say, and yet they have many *different* feels.

I reject this objection, pointing out that there is every reason to believe that sharp pains and searing pains, and so on, have *different* characteristic behavioural profiles (while, of course, sharing the broadly aversive profile characteristic of all pain). A sharp pain, for instance, is a response (at least as if) to something sharp entering the body surface; a searing pain is a response (at least as if) to diffuse surface damage caused by heat. Therefore, on the present account, we can't swap these feels. If we did, an agent with a searing pain would suddenly be motivated to remove an illusory sharp thing, and an agent with a sharp pain would be motivated to mitigate or prevent illusory diffuse surface damage.

Some readers will have noticed that I am using formulations reminiscent of traditional adverbialism in describing my account of qualia. I include a section clarifying that there are several reasons why the analysis offered here is not traditional adverbialism: my position rejects aspects of the sense-data theory which adverbialism still accepted; it goes beyond traditional adverbialism in saying considerably more about what 'sensing redly' (say) involves; and finally, it allows that there really are qualia and that we really do introspect them (where these are more than just linguistic formulations requiring translation to an adverbial format before their truth can be evaluated).

[13] Which empirically can occur, under the influence of strong opiates for instance (Aydede, 2005/2008).

I close by observing that this account in many ways completely reverses the traditional explanatory role for qualia. On the traditional account, we are most directly acquainted with the properties of our sensations (even though, of course, naïve subjects do not take things to be thus). This acquaintance is then used to explain our acquaintance with the world. On the present account, we know all and only the world. Qualia are a public part of the public world (that is, an in principle behaviourally detectable part); any explanation of how we know our own qualia is, on this account, certainly *no more fundamental than* an explanation of how we know the world.

1.2.5 Chapter 6 – Reclaiming Qualia

In the scene setting of Chapter 2, I suggested that we would have enough material for it to count as a plausible naturalisation of qualia, if only we could find introspectible properties which can vary even as between agents who are seeing the same aspects of the world *as* the same aspects of the world, and who can agree that they are doing so.

In this chapter I will claim that over the course of the thesis we have accumulated enough material to naturalise plausibly several other allegedly non-naturalisable aspects of the traditional conception of qualia (in particular, and at least in limited senses, their being knowable *infallibly* and *incorrigibly*, and their *privacy* and *intrinsicness*). I also present one new line of argument which in context shows that a certain form of *ineffability* can be naturalised, too.

I should clarify that I very much agree with Dennett (1988) that the majority of attempts to formalise the above intuitions have ended up as definitions of properties which nothing real could have (and which, therefore, qualia do not have). But I do not agree with Dennett that there are no qualia, or that qualia are only autobiographical fictions (Dennett, 1991). In this chapter I argue that a much better job can be done, than Dennett claims can be done, of naturalising the intuitions which led to these definitions; I argue that we do not need to "get a new kite string" (Dennett, 1991 p.369).

As regards *infallibility* and *incorrigibility*, I draw heavily on Shoemaker's defence of a "limited Cartesianism" (Shoemaker, 1988; Shoemaker, 1990). Specifically, on the fact that one cannot be *rational and wrong* in self-ascription of any mental states which can be fully defined in terms of their role in a space of reasons. Since any such states are *defined* by their role in rationality, and since we cannot be rational and wrong in self-ascription of them, I argue that there is a very good sense in which it is *of the nature* of such states to be known infallibly (we know when we have them, at least if we turn our

mind to it) and incorrigibly (if we think we have them, then we do), for all that these are only ideals which will certainly not always be met in real agents (which are bound to have flawed and incomplete rationality).

Next, I move to *ineffability*. This is the only problematic property of qualia for which I introduce a fundamentally new argument in this chapter. I do so by way of responding to Dennett, once again, but in this case responding to his most recent take on the knowledge argument[14].

Most physicalists, it seems, have agreed that there is no threat to physicalism in Mary's learning something in *some* sense – perhaps gaining an ability (Nemirow, 1980; Lewis, 1983); or learning an old fact, but in a new guise (Churchland, 1985) – on her release. This consensus has recently expanded to include Jackson himself (Jackson, 1998b); Jackson still thinks that Mary will learn something (in some sense), but he no longer thinks that this is a threat to physicalism.

Nevertheless, Dennett still believes that there *is* a threat to physicalism in accepting that Mary learns something, and he has recently (Dennett, 2005b) tried to explain in more detail why.

It might be thought that Dennett *has* to claim that Mary learns nothing. For Dennett is the chief proponent of *heterophenomenology* (Dennett, 1991), and one of the main tenets of heterophenomenology is that the only valid data for answering questions about what it feels like (both in one's own case, and in the case of others) consists in data about what one will say, and how one will react.

As Dennett puts it, at one point in his paper:

> "[It is often supposed that there is] a distinction … between knowing "*what one would say and how one would react*" and knowing "what it is like". If there is such a distinction, it has not yet been articulated and defended, by [anyone] … , so far as I know" (Dennett, 2005b footnote 3).

It might be thought that no such distinction *could* be defended, consistent with heterophenomenology. That much is certainly what Dennett argues. I argue that Dennett

[14] This is Frank Jackson's knowledge argument (Jackson, 1982; Jackson, 1986), wherein Mary is a super-intelligent neuroscientist, who knows and understands everything which science can write down about colour vision, but who has been raised, herself, in a black and white environment. At issue is this question: will Mary learn something, about what it is like to see in colour, on her eventual exposure to the world of colours?

is wrong: that there is such a distinction, even on a strictly functionalist, strictly heterophenomenological account.

The work of Dennett's to which I respond introduces a robot agent, RoboMary. Dennett aims to show how it is that RoboMary would always be able to use her great knowledge to put herself into the state where she does know what it is like. He discusses many variants of, and objections to, his argument, trying to show that some route to knowing what it is like is always going to be available to RoboMary (and hence – I do not wish to question the force of the hence – to Mary).

I argue that a mistake has been made here. I agree that there is nothing unphysical about what any of Dennett's various RoboMary models do. But what should be in question, I argue, is whether a robot which knows as much as Mary *must* be able to come to know what it is like, simply in virtue of knowing as much as Mary (and of having the potential to come to know what it is like, at least on exposure to colour). That is, are the premises of the knowledge argument *alone* sufficient to ensure that Mary must be able to do the kind of thing which Dennett's RoboMary does?

To answer this question, I introduce RoboDennett: a robot who knows as much as Mary and RoboMary, but who is defined to be cheating, for the purposes of the argument, if he uses any ability which is not granted to him simply by the premises of the knowledge argument. I present arguments to the effect that such a robot *does not* have a route to coming to know what it is like, even if it does *know exactly what knowing what it is like consists in*.

Crucially, as I must if I am to take Dennett's heterophenomenology seriously, I explain why there will always be behavioural differences between a robot which knows what it is like, and one which only knows what knowing what it is like consists in. This conclusion is entailed by the account of knowing what it is like given herein; but I show that it is entailed by weaker, highly plausible, engineering considerations about what any state characterisable as knowing what it is like ought to involve.

Having presented these arguments concerning RoboDennett, I can return to the main theme of this chapter. For if the above arguments are correct, then there is indeed a certain *ineffability* to qualia, a certain sense in which they cannot be put into words. For, it turns out, you really do have to experience them (or something logically equivalent, for the purposes of the knowledge argument) in order to know what it is like. Conversely, you cannot put into words what it is like in at least this sense: no mere

description, however well phrased, and however well subsequently understood, will be *sufficient* to make the recipient know what it is like.

Finally, I briefly consider *intrinsicness* and *privacy*. I note, first, that it would certainly go against everything I have argued for, throughout the thesis, to allow either of these in an over-strong sense. Nevertheless, we are talking here about whether it is possible to naturalise underlying intuitions. As such, I draw attention to what I have been claiming (at least implicitly) throughout, that the intrinsicness claim was introduced in the first place, in order to formalise the intuition that qualia can vary in the way described in the inverted spectrum. We have already naturalised that intuition, or so I have argued, and here I argue that by the same measure we have naturalised the intrinsicness claim about qualia.

In the same vein, I certainly do not want private qualia, in any sense which would mean that they cannot by known in others, even in principle; or that they can only be known in others, by comparison with fundamentally first-person knowledge from one's own case; or that they can only be referred to in some kind of private language[15]. Despite all of that, I think we can get a technically 'weak' (but important) naturalisation of privacy.

I show what I mean here by arguing that the present account, despite its reliance on publicly accessible behaviour, can avoid succumbing to the attack on classical behaviourism summed up in the anti-behaviourist joke with the punchline: "it was wonderful for you, darling, but how was it for me?".

Firstly, I observe that one can indeed know all of the properties which I have talked about through introspection. But this is not quite enough. We need the further point which I have already argued for in my Chapter 3 response to Kind, that introspection on the account given here is very much *not* the same thing as the using third person evidence about oneself. As such, the introspectible states of the account I have endorsed (including qualia) are much more personal, more private, than they would be according to this anti-behaviourist joke. I can know my own mental states *just* by turning my attention to them. This need involve no overt sign that I have done so. You, however, can only know my mental states by questioning me, be probing, by finding out what I

[15] To quote Wittgenstein's terminology for the formulation of this problematic conception which he argued against (Wittgenstein, 1953/2001).

will say and do. To put it at its most basic, they are *my* mental states not yours, because I can introspect them and you can't[16, 17].

In sum, I have argued that we can naturalise *infallibility*, *incorrigibility* and *ineffability* relatively strongly, and *intrinsicness* and *privacy* quite well enough to see why people might ever have said such things. This, I suggest, is enough to have reclaimed qualia from Dennett's repeated attempts to quine them.

1.2.6 Appendix – Noë on Experience

In an appendix, I review Noë's recent causal analysis of perceptual experience (Noë, 2003). I include this material since it relates to the thesis proper in several ways. Firstly, I mention this analysis in the main thesis, by way of emphasizing what Shoemaker has lost, if he rules out a causal account of our self-acquaintance with our qualia. Secondly, I defer to this account of Noë's, once or twice in the thesis, as the correct account of 'the right way' for a public object or property to enter an agent's space of reasons, in order for it to count as a *bona fide* case of perception. Finally, this discussion in the Appendix plays one further role, for it allows me to say just a little about issues to do with the conceptual and (on some accounts) nonconceptual contents of perception. These issues have certainly become important to me during the course of this research (I do mention the debate briefly at a few points during the thesis), and I have produced an amount of written work on this topic, though not yet any which has been published nor (for reasons of focus and space) included in the thesis. As such these issues would otherwise remain purely in 'Future Work'.

Briefly, then: I review Noë's recent causal account of perception. I offer a formalisation of Noë's account, of a type which Noë himself gives for the old account which he argues against, but never gives for his own proposed replacement. In passing, I note that a worry which Noë himself offers, as to whether the terminology of his

---

[16] Of course, states can be a creature's mental states even if it cannot introspect them: states are the mental states of an agent if they are the states characterising that creature's occupation of (a part of) the space of reasons (Hurley, 2003). On the account of introspection offered here, introspectibility is a derivative (or, at least, a no more fundamental) criterion.

[17] None of this should be read as denying the truth of the claim that many kinds of mental states (happiness, for instance) are fully and noninferentially visible, right there in the behaviour, in their most paradigm cases (McDowell, 1982). Indeed, at least some mental states, at least some of the time, *must* be manifest in behaviour, in this way, in order for the nonreductive, public-behavioural account of the mental which I endorse here to get off the ground.

account is correct for the case of touch, can be fully dismissed. Finally, although I believe that Noë's account is a major step forward, I argue that it suffers from a notable flaw, in its own terms. Noë presupposes that there is a univocal sense in which we are related to what he calls *factual content* and to what he calls *perspectival content*. I argue that there are analytic reasons for believing that the relevant relationships in the two cases cannot be the same. I argue that Noë's account requires some small amount of sympathetic modification to allow for this issue, and I present the relevant modifications. I argue that the account gains something and loses nothing in the modification.

1.3 Original Contributions

1.3.1 Chapter 2

In Chapter 2, a novel framework is proposed, under which we allow ourselves to be guided in our search for qualia by our best independently plausible theory of introspection. I also propose a novel definition of qualia (or perhaps a better phraseology would be: a novel analysis of one central aspect of the term 'qualia'): this definition requires that qualia are introspectible, but other than that, it is as neutral as it is possible to be about which property qualia are, consistent with naturalising key aspects of the inverted spectrum intuition. However, in this chapter, I also argue *against* the possibility of naturalising even the logical possibility of full, behaviourally undetectable, inverted spectra. In this latter context, I offer a novel line of argument against the modern 'phenomenal concept strategy' (this latter argument is closely related to arguments already given by Chalmers: I present a form of argument directly in terms of explicability which Chalmers mentions as a possibility, but does not develop in the same way, in his own paper on this topic). Finally, I offer suggestive arguments to the effect that many current and historical approaches to naturalising conscious *perception* may have smuggled in (non-naturalisable) theoretical commitments about the nature of *introspection*.

1.3.2 Chapter 3

In Chapter 3, I present the rationality model of introspection, arguing that it is an account shared by Shoemaker and Sellars (although Shoemaker does not, so far as I am aware, anywhere credit his detailed endorsement of this model to Sellars' original, though much less detailed, work on the topic). I present a novel resolution of what at

first appears to be an explicit disagreement between Shoemaker and Sellars about the nature of introspection. The resolution of this only-apparent disagreement helps to clarify what this model of introspection says (and, equally, what it does not try to say) about the physical instantiation of introspective abilities in any given agent. I then present a novel defence of the rationality model, as against recent counter-arguments by Kind. In this context, I present a novel line of argument to the effect that the rationality model has a *better* claim to count as *bona fide* introspection than does the quasi-perceptual model against which it is normally (and in Kind's case) pitted. I also present the novel claim that Shoemaker's arguments in favour of the rationality model of introspection can be generalised, to show that *any* aspect of a space of reasons as such is introspectible (that is, is the right kind of thing to be introspected, on such a model, by some possible agent).

1.3.3 Chapter 4

In Chapter 4 I present Shoemaker's most recent account of qualia. I then present novel criticisms of this account, arguing that it is not compatible with low-level causal explanation (Shoemaker accepts this, though he mentions it only briefly in a footnote), and most importantly, that it is not compatible with personal-level causal explanation. I emphasize quite how high a cost this latter is for Shoemaker, since it calls into question the very analytic-functional understanding of mind within which his account of qualia is supposed to be framed.

1.3.4 Chapter 5

In Chapter 5, I present the central novel analysis of the thesis. I identify certain properties of a space of reasons as such which, I claim, are qualia: I make this claim in virtue of these properties being introspectible (on an account of introspection which, I have argued in Chapter 3, has strong independent plausibility) and being 'subjective' (in the minimal sense identified as being sufficient to naturalise qualia in Chapter 2). I locate these properties as lying within the domains of affect (i.e. motivation) and of learnt and innate association. Both of these features, I argue, are ineliminable elements of a space of reasons as such. These are elements which we do *not* need to specify, if we only wish to specify enough to show that some agent has some mental relation to (or as if to) some public state of affairs. However, crucially, these are elements which we *must* specify, if we wish to move from merely specifying *what* a subject is sensitive to, to specifying what that subject is going to *do* about it. That is, these elements are required

in order to explicitly specify a space of reasons for *action*. In this context, I also briefly present the novel claim that the most basic desire-like state (affect) is not a propositional attitude state, even though the most basic belief-like state (perception) is. I relate this claim to recent work in animal ethology. I present a novel analysis of the various ways which we have of thinking about qualia, and about the effect which public properties (such as objective colours) have, of producing qualia in us. This latter analysis is inspired by elements of Shoemaker's current model of qualia, nevertheless it has the difference, and the advantage, of explaining presentation of a property to a subject in terms of the subject's practical understanding of that property, rather than in terms of some yet-to-be-analysed (and, at least in Shoemaker's case, for reasons given in Chapters 2 and 4, non-naturalisable) representation-relation. I extend the novel analysis of qualitative feel to the case of pain. The central claims are 1) that *pains* (as things perceived; although not *pain*, the feeling), should be identified with (at least seeming) body parts presented painfully (I clarify what this means in behavioural terms), and 2) that the different feels of pain (searing, dull, sharp, etc.) can be accounted for in terms of differences in what damage seems to be present and differences in what the subject is motivated to do about it. Both these claims about pain exist elsewhere in the literature, though of course they occur here in the context of a novel analysis of qualia more generally. I show that the present account is not a reformulation of traditional adverbialism, and is not subject to the (strong) arguments against traditional adverbialism. In showing this, I claim that the account is a form of direct realism. I briefly try to say enough to show why this should not be considered 'threatening': direct realism is a thesis about when *mental* explanation stops, not a thesis about the possibility, or otherwise, of further scientific explanation. I argue that the account of qualia which I offer is also novel within a direct realist context, and explain why this should be so.

1.3.5 Chapter 6

In Chapter 6, I argue that the combination of the present analysis of qualia with Shoemaker's analysis of introspection is sufficient to naturalise our intuitions to the effect that qualia are knowable *infallibly* and *incorrigibly*: it is of their nature to be known thus, just as it is of the nature of belief and desire to participate in rational transitions (and as in this latter case, flaws of and limitations to rationality remain inevitable). This claim is novel as regards qualia (since Shoemaker does not analyse

qualia as fully rational states). I then present novel arguments against Dennett's most recent position paper on the knowledge argument. Dennett believes (despite there having historically been a strong physicalist consensus against this) that there remains a threat to physicalism in accepting that Mary learns something on her release (in *any* sense of 'know' or 'learn'). I adapt Dennett's robot-based style of argumentation in order to show that he is wrong about this, even from the point of view of his own strictly heterophenomenological, functionalist account. I further argue that this result clarifies a certain sense in which qualia are *ineffable*: no formulation in words (however well expressed, and then however well understood) can be sufficient to make the understander 'know what it is like' in the sense at issue in the knowledge argument (and this remains so, even on a strictly physicalist account). Next, I argue that qualia are (weakly) *intrinsic* to the extent that they can (as I have argued throughout the thesis) explain the intuitions which lead to the (over-strong, behaviourally undetectable) inverted spectrum claims and to the formalisation of such claims in the standard (over-strong) intrinsicness claim about qualia. This is a novel line of argument which has been developed over the course of the thesis. Finally, I draw on a point made in more detail in Chapter 3 to argue that the qualia which I have identified are indeed (weakly, but importantly) *private*, as are all our mental states on the rationality model of introspection. This claim relies on the point (which is Shoemaker's originally, but for which I have provided a novel line of defence in Chapter 3) that introspection on the rationality model is a *fundamentally first-person* way of accessing my mental states. Mental states on this account can certainly remain covert and (weakly) private: for, in introspection, *I* do not need to access my behaviour in order to access my mental states; whereas *you* always do need to access my behaviour, in order to access my mental states – including my qualia.

1.3.6 Appendix

The Appendix (whilst not required for the main line of argument in the thesis) also presents a novel contribution, in the form of a novel sympathetic modification to Noë's recent causal analysis of perception. This modification draws on (and clarifies some of the issues within) the ongoing debate in philosophy of mind concerning the conceptual and (allegedly) nonconceptual contents of experience.

2. Background Issues

2.1 Introduction

The purpose of this chapter is scene-setting. In Section 2.2, I present and motivate the approach which I will take in the rest of the thesis, of using an independently plausible theory of introspection to guide the naturalisation of qualia. In Section 2.3 I present the Sellarsian notion of mind as physical locus of action for reasons, on which many aspects of the thesis are founded.

2.2 Qualia and Introspection[18]

2.2.1 Abstract for this Section

The claim that behaviourally undetectable inverted spectra are possible has been endorsed by many physicalists. I explain why this starting point rules out standard forms of scientific explanation for qualia. The modern 'phenomenal concept strategy' is an updated way of defending problematic intuitions like these, but I show that it cannot help to recover standard scientific explanation. I argue that Chalmers is right: we should accept the falsity of physicalism if we accept this problematic starting point. I further argue that accepting this starting point amounts to at least implicitly endorsing certain theoretical claims about the nature of introspection. I therefore suggest that we allow ourselves to be guided, in our quest to understand qualia, by whatever independently plausible theories of introspection we have. I propose that we adopt a more moderate definition of qualia, as those introspectible properties which cannot be fully specified simply by specifying the non-controversially introspectible 'propositional attitude' mental states[19] (including seeing *x*, experiencing *x*, and so on, where *x* is a specification of a potentially public state of affairs). Qualia thus defined may well fit plausible,

[18] Section 2.1, *Qualia and Introspection*, is forthcoming as a paper in the *Journal of Consciousness Studies* (Beaton, in press) with only very minor differences made to the version given here, as required to link the work to the rest of the thesis.

[19] Throughout this thesis, 'state' will be used to mean 'state or process'. This usage of 'state' is ubiquitous in the physical sciences, where a physical 'state' can easily be characterised in such a way that a physical system in that instantaneous state *must* be in different instantaneous states at different times (i.e. by characterising the state as an instant in a time-varying process).

naturalisable accounts of introspection. If so, such accounts have the potential to explain, rather than explain away, the problematic intuitions discussed earlier; an approach that should allow integration of our understanding of qualia with the rest of science.

2.2.2 Overview

In this section I will be concerned, in a certain sense, with the *definition* of consciousness; that is, I will be discussing the nature of the *target* of explanation in our scientific or philosophical study of consciousness. As many authors have observed (e.g. Rosenthal, 2002; Vimal, in press), there are multiple views about how to pin this target down. Ought we to be trying to explain consciousness conceived of as a cognitive property? As a phenomenal property? As somehow related to awareness and attention?

This thesis is concerned with the phenomenal aspect of consciousness: with *qualia*; with the 'something it is like' to have an experience[20]. This is not to completely ignore the many other aspects present within the broader concept of 'consciousness', as covered by Vimal and others. Indeed, it is my hope that many or most of these aspects will prove to be intimately related to each other, within the right theoretical framework. Nevertheless, there is a certain mystery to the phenomenal aspect of consciousness in particular. It seems especially hard to find a place for that aspect within our growing understanding of the natural world (Chalmers, 1995; Levine, 1983).

The aim here will be to critique a particular approach to phenomenal consciousness which 'defines in', from the start, certain problematic features of qualia. Specifically, I will critique that class of approaches which entail that our knowledge of phenomenal facts is *a posteriori* with respect to our knowledge the physical facts.

There is quite a lot to be unpacked here, about what philosophers mean when they talk like this. To get the discussion started, I need to introduce two assumptions which I share with the position I am critiquing. The first is this: when I introspect and come to think that it is 'like this' for me to see red (for example), then my thought refers to some fact: a fact about 'what it is like' (or, equivalently, about what the phenomenal feel is). We can call such facts *phenomenal facts*, and knowledge of such facts *phenomenal knowledge*. The second shared starting point is this: it is possible to discover the

[20] Qualia are the characteristic properties of phenomenal consciousness: something is a state of phenomenal consciousness if and only if it has such properties.

existence of regular co-occurrence between physical facts and introspectible phenomenal facts. If so, we would be able to discover that when certain physical facts about a creature ('neural correlates of consciousness') or, perhaps better, about a creature in its world (physical correlates of extended mind), are true, then certain phenomenal facts are always true.

Given these shared starting points, the *a posteriori* approach which I am critiquing goes on to claim that the existence of this regular co-occurrence between public physical facts and introspectible phenomenal facts could not have been worked out in advance, purely by conceptual analysis, however well we understand what we mean, when we say that we 'know what it is like' and however well we understand the public physical facts which co-occur with the phenomenal facts.

David Chalmers has called this kind of approach *phenomenal realism* (Chalmers, 2003a). As Chalmers rightly states (2003a), and as I will show below, certain very common presuppositions about phenomenal facts (specifically, either or both of the inverted spectrum[21] or zombie[22] claims about qualia) directly entail that there is this kind of *a posteriori* relation between physical and phenomenal facts. Chalmers also states that it is not possible to "take consciousness seriously" (Chalmers, 1996 p.xii), without adopting starting points which lead directly to such a view. For the purposes of the present work, I will use the label *strong phenomenal realism* for such views, since my main aim will be to claim that there are other ways to take qualia seriously.

The biggest problem with such *a posteriori* approaches is that they rule out (on the basis of presuppositions built into their definition of qualia) a certain extremely standard form of scientific explanation. In Section 2.2.3, I will outline the model of explanation in question. Then, in Section 2.2.4, I will present one historically popular (and still influential) approach to naturalising qualia which I will use as an example, to make clear why these starting points rule out this type of explanation. In Section 2.2.5, I will outline the modern phenomenal concept strategy, which claims that physicalism can be preserved, even if we adopt such *a posteriori* claims about qualia. I will then present

[21] The claim that there can be creatures which are physically (or functionally) identical to each other, but which have different phenomenal mental lives.

[22] The claim that there can be creatures which are physically (or functionally) just like us, but with no phenomenal mental lives at all.

reasons to agree with Chalmers, when he claims that this phenomenal concept strategy cannot work.

The final parts of Section 2.2 question whether theorists really are entitled to such problematic starting assumptions. In Section 2.2.6, I will argue that such starting points amount to implicit *theoretical* claims about the nature of introspection: claims which, if true, are themselves justified by introspection. I will point out that there is widespread disagreement about the nature of introspection, and I will suggest that there is a widespread tendency to build presuppositions about it into our theories of sensory experience. As such, I will argue that the theorists I am critiquing are not justified in endorsing such problematic starting points.

Finally, in Section 2.2.7, I argue that it is possible to preserve a moderate form of phenomenal realism (there really are qualia, we really do know them in introspection), without these problematic starting points. To do this, I propose a more moderate definition of qualia, which allows our theorising about them to be guided by whatever independently plausible theory of introspection we have. I will argue that this moderate definition still looks to have the ability to explain, rather than completely explain away, many intuitions about qualia, including some of the problematic starting points above.

2.2.3 Normal Scientific Explanation

In this section I will briefly present an account of a very standard form of scientific explanation. My claim is that this form of explanation is so ubiquitous, that for any property which science recognises, the existence of that property is either a) believed to be explicable in terms of more fundamental properties in this way, or b) is treated as a fundamental fact about our universe.

A paradigm example is the explanation of the properties of water (the way it freezes and boils, its transparency, its viscosity, and so on) in terms of the properties of, and interactions between, H_2O molecules (the shape of the molecule, the forming and breaking of hydrogen bonds between molecules, and so on).

Philosophers often like to emphasize the fact that the relation between water and H_2O molecules can only be known *a posteriori*: that the existence of such a relation could not have been worked out in advance of the relevant empirical discovery, even with the most careful reasoning. But this is a misdescription, or at least an over-simplification. As Loar (1997 p.608) and Chalmers (2006), amongst others, have noted, there is an *a*

priori entailment between the low level properties of H_2O and the high-level properties of water.

It is possible to be too prescriptive about exactly what such an *a priori* entailment involves (see note 23), so I will try to put it as neutrally as possible: having mastered the concepts involved in describing the low and high levels, it would not be rational to believe that certain high level facts do *not* apply (e.g. that there is stuff which behaves like water round here) when certain low level facts apply (that there is a large number of H_2O molecules with a certain energy distribution, etc., around here). This is an *a priori* conceptual entailment, in that the existence of the rational link in question follows purely from an understanding the concepts involved, with no further empirical research necessary[23].

Note, also, that it is a *one way* conceptual entailment: the facts[24] about H_2O molecules entail that a mass of them behaves the way water behaves, but the facts about the way water behaves do not entail that it is made of a mass of H_2O molecules. I would agree that it is not rational for someone informed by modern science to claim that water is *not* (mainly) made of H_2O. But the logic in this direction is fundamentally *a posteriori*, based on induction from the *discovery* that what has been found to explain wateriness round here always has been H_20. This relationship between concepts, which is *a priori* in one direction but *a posteriori* in the other, can be contrasted with *two way* cases such as the relationship between 'bachelor' and 'unmarried male' (*a priori* in both directions), or that between 'son of Barack Obama, Senior' and '44th president of the

[23] In fact, this is not an *a priori* entailment in the strict philosophical sense: a step requiring *no* empirical knowledge whatsoever. This is because the kind of practical mastery of the concepts required to see the connection between the high and low levels *does* require empirical knowledge and experience. The account I'm giving therefore claims that we use common sense, at the point where the more traditional 'deductive nomological' account of scientific explanation/reduction would claim that we use 'bridge laws'; but I don't think anything in the main line of argument hinges on this difference from the perhaps more familiar account. For these and various other reasons, the account I am giving is not quite that of Chalmers and Jackson (2001).

[24] A note on how I individuate facts in this thesis: I treat the fact that 'H_2O molecules are present' as a different fact from the fact that 'water is present' (even when they refer to one and the same state of affairs), because of the (one-way) conceptual independence between the levels of description involved; conversely, I would treat the fact that 'a bachelor is present' and the fact that 'an unmarried male is present' (when they refer to the same state of affairs) as the same fact, because there is no conceptual independence between the two descriptions involved.

United States of America' (*a posteriori* in both directions). I am claiming that the special, one way kind of relationship is essential for scientific explanation[25]. It is important to be clear that the high level properties do not somehow disappear once we have such an explanation: it is only in talking at the high level that we can express what needed to be explained in the first place. In fact, the concepts of the high level need not even be applicable at the low level.

This pattern is not specific to water and H_2O; it is widely repeated, in scientific explanation. The same pattern holds between the micro-facts of modern genetic theory (transmission of DNA, gene-expression during embryonic development, etc.) and the macro-facts of inheritance with variation required for Darwinian evolution[26], or between the micro-facts of statistical mechanics and the macro-facts of thermodynamics, and so on and so on.

Unfortunately, many views which take qualia seriously, including many which see themselves as varieties of physicalism, build elements into their definition of qualia which rule out any chance of providing explanations of this type.

2.2.4 The Nature of Functionalism

There is a historically popular brand of functionalism which tries to argue that inverted spectra are perfectly possible, and are compatible with normal science. The view was advocated (with subtle differences, on which see more below) by Lewis (1980), the Churchlands (1982) and Shoemaker (1975), amongst others. Lewis says:

> "As philosophers, we would like to characterize pain a priori. … As materialists, we want to characterize pain as a physical phenomenon." (Lewis, 1980 p.123)

An *a priori* characterisation of pain would be one which makes clear that certain facts (e.g. wincing, groaning, withdrawing from noxious stimuli[27], etc.) are two way conceptually identical to facts about pain. Such an *a priori* characterisation of pain would presumably be just a small part of an *a priori* characterisation of the entire

[25] See Section 2.2.4.1 for a brief discussion of an opposing view.

[26] As in many such cases, we have enough of the detail so that the relation between the levels no longer seems 'in principle' mysterious – even though many details remain to be discovered, and our understanding of both levels may no doubt be refined in the process.

[27] Or, at least, a tendency towards such behaviours, which may be masked by other factors but which could be revealed by suitable experimentation.

mental level (including belief, desire, perception and so on) applicable to any agent with a mental life.

It is a general characteristic of functionalism (not just of the particular variant being discussed here) that it supposes that there exists some level of characterisation of a creature which is 'the mental level', and that there are other facts about that creature which can vary, independently of the mental level. This seems to me to be the right kind of approach (with caveats about exactly how this approach should be understood, which I will explain below). In the case of the type of functionalism I am discussing here, however, this strategy is *not* followed through to what might seem its logical conclusion. For the *a priori* characterisation of the mental level is supposed, by these authors, *not* to capture everything mental which there is to say about the subject. Specifically, it does not capture what it is like to be the subject; it is supposed that there could be two subjects who are the same, in terms of this publicly observable mental level of behaviour, but where it nevertheless feels one way to be one subject, and another way to be the other.

In making this point, the Churchlands mention the classic inverted spectrum case, in which we are asked:

> "to imagine someone … [who has] a sensation of red in all and only those circumstances where you have a sensation of green, and so forth." (Churchland and Churchland, 1982 p.122)

The Churchlands explicitly claim that:

> "These cases are indeed imaginable, and the connection between quale and functional syndrome is indeed a contingent one." (Churchland and Churchland, 1982 p.122)

In a similar vein, Lewis asks us to:

> "Suppose that the state that plays the role of pain for us plays instead the role of thirst for a small subpopulation of mankind, and vice versa." (Lewis, 1980 p.128)

Lewis argues that in such a case:

> "there is no determinate fact of the matter about whether the victim of the interchange undergoes pain or thirst." (Lewis, 1980 p.128)

This claim would be false if the phenomenal feel were fully determined by the functional role: if so, a groaning, writhing[28] agent would be unequivocally in pain,

[28] On any plausible *a priori* account of the mental, it must be supposed that the groaning and writhing is suitably integrated with other aspects of the agent's behaviour, quite possibly including their rationality.

whatever was the case about the physical states constituting the agent. But the authors quoted here think that there are two meanings of pain, the *a priori* meaning, where pain simply refers to that state where a creature displays (or tends to display) pain behaviour, and the *a posteriori* meaning, which refers to whatever physical state science has determined to fill this functional role (in a population) [29].

I will describe such views as *hybrid functionalism* (c.f. Lewis, 1980 p.124), since they combine elements of the earlier identity theory ('the physical stuff determines the feel') with what would otherwise be 'pure' functionalism (the claim that the mental facts are fully captured at the in principle publicly observable mental level).

Why, though, believe that a difference in "physical realization" has any "bearing on" the introspectible facts about "how that state feels"? (The quotes are from Lewis, 1980 p.130.) The Churchlands flesh out this part of the view in more detail:

> "the spiking frequency of the impulses in a certain neural pathway need not prompt the non-inferential belief, "My pain has a searing quality." But withal, the property you opaquely distinguish as "searingness" may be precisely the property of having 60 Hz as a spiking frequency." (Churchland and Churchland, 1982 p.128)

The claim is that the physical state of 60 Hz neural firing (or whatever physical state it really turns out to be) *is* what we introspect, when we introspect a searing pain. Equally, in some other agent, the same functional role might be filled by a different physical state, such as inflation in hydraulic cavities in the feet (Lewis' semi-humorous suggestion as to the state which might play the role of pain in Martians). A difference like this is supposed to be the right kind of difference to account for a difference in introspectible feel, of the kind involved in the inverted spectrum (see also Shoemaker, 1975, e.g. p.310).

There is a problem with such views, though, if we want to look for a scientific explanation of qualitative feel, of the form already outlined in Section 2.2.3. It is not that there are no low level differences with which to explain the alleged difference in feel; as we have just seen, there are. The problem is that there would seem to be no high level difference at all, in the central case of behaviourally undetectable inverted spectra.

This is a point which both the Churchlands (1982 p.128) and Shoemaker (1990 p.71) make. See also Section 5.4.3 for further comments on the relevance of this point to the present work.

[29] There are issues here, to do with whether, and in what sense, sub-system states could possibly be role fillers for mental level states such as pain (see, e.g. Shoemaker, 1990 p.67). I won't say much about this at this point, although I will say much more in Chapters 3, 4 and 5 (see also footnote 46 in this chapter).

For two such creatures will do and say exactly the same things. Each will say "it feels like this". If you ask them *how* it feels, they will say all the same things as each other (e.g., "it feels searing"). And so on, and so on. The above model of scientific explanation can only work if we have differences at the low *and* the high levels (e.g. certain stable hydrogen bonds are formed; water freezes). With no difference at the publicly observable mental level, we are left looking for a reason to suppose that there is any mental level difference at all. It is at this point that the various authors mentioned differ.

2.2.4.1 Explanation and Reduction

We only have a problem, as regards giving an explanation of the type outlined in Section 2.2.3, if there are indeed two different levels to relate: a level of mental facts (which do not entail any lower level, non-mental facts), and some non-mental facts (whose existence is not entailed by the mental facts, but which might – if a standard explanation can be given – entail those facts). As we have seen, this is no more nor less than is the case with water versus H_2O, or with heat and temperature versus statistical distribution of energy across microstates. However, in the case of the mental, the existence of such a conceptually separate higher level can be denied.

To see what would be involved in this denial, we need to notice that there are two different ways of understanding the proposal that we should look for an *a priori* analysis of the mental, only one of which I would endorse. I endorse the claim that there is an *a priori* relation between the public notion of pain, and a tendency towards certain behaviours such as wincing, groaning, withdrawing from painful stimuli, etc. But I am endorsing this as a relation amongst facts *at the same level*. Thus pain, wincing, groaning, etc. are all (in the first instance) *mental* level facts[30], just as the properties of macroscopic water (boiling, melting, etc., etc.) are all 'water level' facts.

There is an entirely different reading of the same claim which I would *not* endorse. On this reading, wincing, groaning, withdrawing, etc. are to be read as entirely non-mental facts, and the claim being made, in that case, is that the mental level is not conceptually independent of such entirely non-mental facts. If it is right that the mental is identical (on careful reflection) to some non-mental level of description, then it could

[30] At least, wincing and groaning are mental facts, to the extent that they occur with the right connections to the rest of the mental – see note 28.

be coherently claimed that introspecting the feel of pain is conceptually the same thing as subpersonal detection of a subpersonal state such as 60 Hz neural firing (when this occurs within the right, surrounding subpersonal context).

If this fully 'operationalized' *a priori* analysis of the mental can be carried out, then we don't need to look for an explanatory relation between two levels of description (as outlined in Section 2.2.3), because there is really only one level of understanding in play.

Endorsement of this latter kind of *a priori* analysis is a very strong form of reductionism about the mental (which is sometimes not clearly enough distinguished from the process of *explanation* outlined in Section 2.2.3). In many ways, this strongly reductive approach looks like a denial of the reality of the mental level[31], especially when it is made clear that no such conceptual reduction is involved in the explanation of many much less contentious properties[32]. As such, in the rest of Section 2.1, and the rest of the thesis, I will discuss what follows if we assume that there *is* a conceptually separate mental level, and that what we are looking for is an *explanatory* relationship between non-mental facts and mental facts (or, at least, an understanding of why we cannot have such an explanatory relationship). On this, at least, I agree with Chalmers, with the authors working on the phenomenal concept strategy (Section 2.2.5) and with a least one of the authors who historically argued for hybrid functionalism.

2.2.4.2 Phenomenal Knowledge

The above strongly reductive analysis would indeed give us a reason to believe in a mental difference between some functionally identical agents: if the analysis is correct, a physical difference of the right type *is* a mental difference. However, if we don't accept the reductive analysis, then we still have no third-person reason to believe that

---

[31] One might call such an approach *eliminative reduction*, but it is *not* the same thing as the outright *eliminativism* which the Churchlands argued for elsewhere, concerning the belief-desire framework of folk psychology (see, for instance, the sections on eliminativism in Churchland and Churchland, 1998); one cannot hope to show that 'introspecting phenomenal feel' is conceptually identical to some reasonably well-defined set of subpersonal processes, if one also wishes to show that 'introspecting phenomenal feel' is part of a bad conceptual scheme which does not refer very well to anything at all.

[32] In fairness to the Churchlands' position, I should make clear that they did not accept the analysis of scientific explanation which I have given. Instead they asserted that the pattern of conceptual analysis of role, coupled with *a posteriori* discovery about role filler, is normal elsewhere in science (Churchland and Churchland, 1990, e.g. p.78).

there is a mental difference between the supposed inverts. That this is so follows in two steps. Firstly, there is no reason at the public mental level to suppose that there is a difference, for such agents are *the same* at that level. Secondly, the publicly accessible difference which *does* exist between alleged inverts (on the hybrid-functionalist view; e.g. 60Hz neural firing vs. inflation of hydraulic cavities) is a physical difference: it lies at a level of description which is not (without further argument) mental. Without the reductive claim, and considering *purely* the third-person facts, there is no reason to believe that that public difference is (or causes, or amounts to) a mental difference.

Nevertheless, not all of the authors who have argued for the compatibility of functionalism and the inverted spectrum intuition endorse a strongly reductionist analysis. Shoemaker, for instance, was not and is not a reductionist about the mental, but he took and takes the inverted spectrum intuition seriously as a starting point for theorising about qualia (Shoemaker, 1975; Shoemaker, 1994c; Shoemaker, 1994d). It should be emphasized, then, that it follows logically that, if one endorses the strong phenomenal realist view, but rejects reductionism, one *must* take oneself to have a *first-person* reason to believe that the inverted spectrum is possible[33]. It is worth emphasising clearly what this means. Without reductionism, there can be *no reason* to believe in inverted spectra *at all*, unless it is a reason which fundamentally involves first-person knowledge. If such views are right, we *must* be able to come to know by introspection[34] that 'what it feels like' is the kind of thing which could differ, even as between two agents who act in all the same ways[35].

Now we can see the connection between the strong phenomenal realist starting points (specifically, the zombie or inverted spectrum claims; though we are mainly considering the inverted spectrum claim, since this is the one popular with many physicalists) and *a*

---

[33] I am ignoring the complications which might follow if, for instance, someone claimed that the inverted spectrum intuition was grounded in fundamentally second-person (Thompson, 2001; De Jaegher, forthcoming) knowledge.

[34] I will treat 'introspection' as identical to 'the ability to gain knowledge in a fundamentally first-person way'; even if the relevant knowledge is not gained *entirely* through introspection (in this sense), it must be gained in a way which *essentially* involves introspection.

[35] Informal conversation indicates to me that a large number of (though not all) thoughtful non-philosophers do indeed take themselves to know exactly this; they take themselves to know, presumably on the basis of introspection, that the inverted-spectrum scenario is 'obviously' possible. So this starting intuition, if wrong, is widely (though not universally) shared, at least in this culture.

posteriori knowledge. For the knowledge which one is supposed to have, on such accounts, is knowledge which cannot be entailed by (just) the third-person facts, since none of those facts (taken apart from introspective knowledge) give us any reason to believe that there is a mental difference, as we have seen. Equally, if we are not being reductionist about the mental level, then there is no reason to suppose that the mental facts on their own (including any facts known by introspection) entail any lower level, non-mental facts. So here, we have a pure (i.e. two way) *a posteriori* discovery – there are certain phenomenal facts which I know, when I 'look' inwards (i.e. introspect) which I could not have known by looking outwards[36].

It turns out, then, that the same starting points which entailed that there was no publicly accessible high level to explain (in certain key cases) must also entail that phenomenal knowledge is entirely *a posteriori* with respect to (neither entailing nor entailed by) our knowledge of publicly observable facts[37]. Note that this kind of knowledge is strange in that (if it really exists) its existence is *a posteriori* with respect to (i.e. it could not have been deduced from) all knowledge of the third-person facts, however clear thinking and detailed.

Even with the need for this unusual kind of knowledge, perhaps it might still be argued that these views are not so implausible after all. For whilst this is a very special kind of knowledge (c.f. Chalmers, 1996 p.193), it is also knowledge of a special kind of state. Perhaps we should *expect* ourselves to have non-standard and intimate knowledge

---

[36] The disconnect between this alleged knowledge and knowledge of publicly accessible facts is much *stronger* than the 'disconnect' between public knowledge and indexical knowledge (first-person knowledge such as "I am in Sussex", "It is Sunday", etc.). This is because the fact that I can only have indexical knowledge when I am *in* a certain state *follows* from the publicly observable facts, plus an understanding of the concept of indexical knowledge (see Chalmers and Jackson, 2001; related points are made in Beaton, 2005 and Section 6.4). Whereas the phenomenal knowledge which (allegedly) grounds our belief in the possibility of the inverted spectrum has to be of a quite different type: it might well be possible to *learn* (*a posteriori*) that when I am in a certain physical state, I will be in a certain phenomenal state, but there can be no communicable understanding of the nature of this phenomenal state which could let someone *work out* (*a priori*) that when an agent is in the physical state, the agent *must* be in the related phenomenal state.

[37] Actually, these starting points only strictly rule out an entailment from physical facts to mental facts. There could still be (just) the reverse entailment. This would make (at least some) mental facts more fundamental than any physical facts. This is a form of idealism, and certainly not a rebuttal of the claim that strong phenomenal realism rules out physicalism, which is what I am trying to establish.

of those states which partly constitute us? Indeed, in some sense of this suggestion, I would agree with it. But perhaps it is the case that such intimate knowledge *ought* to have these strange *a posteriori* features? Considerably more would need to be said here, to defend this suggestion. As far as I am aware, the hybrid functionalists whose position was outlined above never said it[38], but more recent work in the philosophy of mind has stepped in to fill the gap.

2.2.5 The Phenomenal Concept Strategy

Loar (1997), and the other proponents of 'the phenomenal concept strategy', embrace the point which I have just made, that what we know about qualia from the first-person is two-way conceptually independent of any facts which science might access. Thus Loar says:

> "Phenomenal concepts are conceptually irreducible in this sense: they neither a priori imply, nor are implied by, physical-functional concepts. Although that is denied by analytical functionalists[39] … , many other physicalists, including me, find it intuitively appealing." (Loar, 1997 p.597)

But Loar also argues that this need not be a problem for physicalism:

> "It is my view that we can have it both ways. We may take the phenomenological intuition at face value, accepting introspective concepts and their conceptual irreducibility, and at the same time take phenomenal qualities to be identical with physical-functional properties of the sort envisaged by contemporary brain science." (Loar, 1997 p.598)

How could such a view work? The general strategy (shared by Loar and others who've published variants of this view) is to concentrate on the special way which we have of introspectively thinking about our own phenomenal states. The claim is that the phenomenal concepts[40] involved in such thoughts ('*this* feeling'; 'like *this*') are special,

---

[38] Of course, for the reasons outlined, the Churchlands needed no such account. For suggestions from Shoemaker along these lines in more recent work, see Shoemaker (1994c Section IV) (a relevant passage is quoted in Section 4.1 of this thesis).

[39] Loar is referring to the thoroughgoing variety of functionalism which takes *everything* mental to be analysable in terms of its (at least counterfactual) relation to publicly accessible behaviour (i.e. he is not referring to the hybrid variety of functionalism I have just been discussing).

[40] Concepts, in the sense used here, do not require language: rather, they are the recombinable components of rational thought. In the same vein, rationality itself, as used here, should be understood in a sense whereby a rational agent is one which can *make* rational decisions, not necessarily one which can

in that they are "conceptually isolated" (Carruthers and Veillet, 2007) from the third-person concepts which we use when we think about publicly accessible facts. The claim that phenomenal concepts are conceptually isolated does not mean that they cannot occur in the same thoughts as publicly applicable concepts. But it does mean that no amount of *reasoning* can lead from facts expressed using phenomenal concepts (e.g. 'my experience is like *this* now') to facts expressed using publicly applicable concepts (e.g. 'my physical-functional state is this, now'), or *vice versa*.

Apart from this general point about conceptual isolation, the views vary as regards the specific nature of phenomenal concepts which is supposed to explain the isolation. Loar (1997) and others have equated phenomenal concepts with some form of recognitional concept; Perry (2001) has equated phenomenal concepts with some form of indexical concept; Papineau (2002) has suggested that phenomenal concepts are 'quotational' ("my red is like this: ____", where the blank is filled in by the experience itself). As such, all these views are trying to give a more detailed account of the first-person *acquaintance* which we have with our own qualia[41] – i.e. an account of exactly what seemed to be missing, in the variant of functionalism outlined above.

Can such a view successfully preserve physicalism? A lot has been written about the phenomenal concept strategy, and I don't wish to dismiss it out of hand. Nevertheless, there is a very general argument against the possibility of phenomenal concepts preserving physicalism[42], if physicalism is understood as requiring an explanation of the presence of consciousness in the manner outlined in Section 2.2.3.

First of all, it is worth noting that the *a posteriori* claim about the nature of phenomenal knowledge (which is so central to the phenomenal concept strategist's

---

make rational decisions by thinking them through, step by step, in the manner of the most complex human thought (c.f. Section 3.3.3.2).

[41] More accurately (c.f. Chalmers, 2003a), an account of the knowledge which such acquaintance can grant us. In the sense in which Chalmers uses the term, the *acquaintance* itself comes in simply having the quale; but this acquaintance is the fundamental ground for later first-person, conceptual knowledge of the quale. It should be noted that a moderate phenomenal realist, type-A (c.f. Chalmers, 1996) materialist (i.e. the position which I am trying to defend, or at least open a space for, in the present work) can, I think, feel very sympathetic to much of what Chalmers (2003a) says about the nature of acquaintance; that is, can feel that very much of it ought to be naturalisable (for more on this, see Section 5.6).

[42] The quick argument given below is very closely related to the central argument towards the same conclusion presented in Chalmers (2006). The main difference is that I proceed directly in terms of explicability, rather than via conceivability.

position) is not merely entailed by the inverted spectrum starting point (as I have already shown, in Section 2.2.4.2), it also entails it. To see why this is so, note that the denial of the inverted spectrum starting point amounts to the claim that there *is* always a behaviourally detectable difference, for every difference in qualia. The notion that there exists special *a posteriori* knowledge of the phenomenal is not compatible with this denial of the inverted spectrum. That is, the phenomenal concept strategist cannot accept an analysis of phenomenal concepts which shows that, for every difference known that way, there must be an observable difference in behaviour at the public mental level[43]. If there were such an analysis, a difference in physics sufficient to explain these publicly observable differences would be sufficient to explain the difference in qualia[44] (on the model of the explanation of the properties of water). The connection between the physics and the phenomenal level would not be *a posteriori*, after all.

It might be thought that the phenomenal concept strategist could still claim that, whilst there can be no conceptually necessary difference in behaviour corresponding simply to a difference in qualia, there still might be a conceptually necessary difference in behaviour corresponding to an agent *knowing* one thing as opposed to another about their own qualia. But actually, they cannot accept this either. Even if qualia are 'covert' when not known about, and only become 'overt' when known about, the normal model of explanation can get a grip. Any physical description which shows why there are these behaviourally observable differences (in the cases where the differences are overt) and why there are no behavioural differences (in the cases where the differences are not overt) will once again explain the physical nature of qualia (on the model of the explanation of water). Once again, the connection between the physics and the phenomenal level would not be *a posteriori*, after all.

I don't think any of this pushes the phenomenal concept strategists to a position which they would be unwilling to accept. It seems very close to (and perhaps actually) explicit in the approach that certain phenomenal differences (and, equally, certain

[43] They could perhaps accept the bizarre position that whilst there is no reason (which we could ever understand) for there to be such a difference in every case, it nevertheless turns out that there is such a difference in every case.

[44] It is important to the argument that I specified that for *every* difference known, there is a (an at least counterfactual) difference in behaviour – this is what the phenomenal concept strategist cannot accept.

differences in phenomenal knowledge) will not result in any behaviourally detectable difference.

The trouble with all this is that it makes quite clear that the phenomenal concept strategy is entirely incompatible with an explanation of the status of qualia along the lines outlined in Section 2.2.3. Not only are qualia themselves not naturalisable along these lines, but the special phenomenal knowledge which is supposed to save physicalism is (and must remain) inexplicable for the very same reasons. We seem to be back to square one[45], with no third-person reason to believe that knowledge of this type exists. Even if we do have a first-person reason to believe this (and Sections 2.2.6 and 2.2.7 argue against that claim), we are left with an unsatisfying, purely 'ontological' physicalism in which we can have no explanation of why certain things are part of the physical world, merely an acceptance that they are.

In fact, I wonder whether things are not worse than this, for the phenomenal concept strategists. Their claim is that the existence of this type of phenomenal knowledge is itself not entailed by anything which physics can teach us (however well we understand the physics, and the concept of phenomenal knowledge). If this is correct, then surely Chalmers (1996) has been right all along? Surely all the physical facts might have been exactly the same, and the phenomenal facts might have been different, or absent altogether? At least, if this is not so, physics can't *explain* why it is not. As such, it looks to me as if Chalmers has been the most honest here, all along. *If* you start from the assumption that there is a pure (i.e. in both directions) *a posteriori* relation between the phenomenal and the physical, or *if* you start from the assumption that behaviourally undetectable inverted spectra are possible, *then* you should end up where Chalmers ends up: you *should* accept that phenomenal properties, and any principles bridging them to normal physical properties, are fundamental facts about our universe.

In the remaining sections of this discussion (2.2.6 and 2.2.7) I want to ask two questions. First, what justifications are there for taking the problematic strong phenomenal realist starting point? Second, if the relevant justifications are found

[45] Actually, as Chalmers notes (2006 Section 4), the phenomenal concept strategy has at least made the genuine contribution of clarifying that strong phenomenal realism entails the existence of this type of knowledge. I would argue (and again, I think most phenomenal concept strategists would be quite happy to agree with me) that the main aim of such accounts must therefore be to convince us that we are wrong to want an explanation of the type I have described, in the case of qualia or of phenomenal knowledge: that physicalism does not require this.

wanting, what could we use as a replacement starting point, if we still want to naturalise qualia?

2.2.6 The Properties of Sensory Experience

Qualia are properties of sensory experience broadly construed to include states such as seeing, hallucination, sensory memory, sensory imagination, and so on. Furthermore, as we have seen above, if there is any reason to believe that qualia are problematic in the way in which the strong phenomenal realist claims they are, this reason must be introspective.

But there is very little agreement about what sensory experience consists in, and even less agreement as to what the introspectible properties of sensory experience are (c.f. Crane, 2005/2008; Gertler, 2003/2008). I know that I am seeing a scarf on the desk in front of me (it is cold round here, right now!); but can I know that I am seeing this in virtue of some more direct kind of acquaintance with sense data? Sense data theorists certainly thought so, but this view is now widely agreed to be false. Can I know that I am seeing the scarf in virtue of, or at least accompanied by, qualia which can vary free of the physical facts? Chalmers and many others have thought so; but many others again *don't* share this certainty. On a related note, the reductionist approach taken by the Churchlands entails that what we know in introspection (of pain states, of colour experience, and so on) includes opaque knowledge of the *physical* nature of certain subpersonal states which underpin these sensory experiences; this, too, is impossible according to many other theories of introspection[46].

Note that all of the above mentioned claims about *perception* (that it involves sense-data; that it entails the possibility of introspective knowledge of the physical states underlying it; that it is accompanied by behaviourally undetectable qualia) constrain our eventual theory of *introspection*, which has to be such as to allow for introspective knowledge of the problematic states in question. Moreover – arguably in all cases, and certainly in the case of the view which is being critiqued here (strong phenomenal

[46] Indeed, this is impossible on any theory in which the facts introspected are all at a conceptually independent mental level, e.g. Sellars (1956), Shoemaker (1996) (for much more detail on these theories, see Chapter 3). This conclusion follows as long as the conceptual independence of the mental level from the physical is at least as strong as (but it need be no stronger than) the conceptual independence of the water-level from the H_2O level.

realism) – whatever plausibility these starting points have itself derives from introspection of such perceptual and experiential states.

Sense-data theorists certainly did take themselves to have introspective knowledge of sense-data. It strikes me as highly plausible that this assumption was an input to the sense-data theory, not an output from it; that the theory made explicit what already seemed introspectively obvious. But, it is widely agreed, the theory was false – we have no such knowledge for there are no sense-data.

Equally, as we have seen, at least some physicalist advocates of the inverted spectrum have taken themselves to have opaque introspective knowledge of the *physical* nature of certain of their internal states. Again, is this input or output? With certain implicit, but theoretical, assumptions about introspection under one's belt, it can seem more or less obvious that we do have introspective knowledge of the physical states which constitute us. But actually, the claim that introspection is like this is a major theoretical assumption. It cannot be justified as a starting point, unless we *already* (i.e. entirely pre-theoretically) have introspectively based knowledge, which entails that it is true. Do we have such knowledge? It seems to me very hard to see how we can decide the case either way, simply by introspecting 'harder' or 'more carefully', and very easy to become misled by one's theoretical commitments.

The same points certainly apply to strong phenomenal realism. As we have seen, the *starting* point of the view is this: there is something which we know by introspection, which is a valid basis for the claim that phenomenal facts cannot be deduced from publicly observable facts[47]. Viewed with some perhaps healthy scepticism, this looks very like an implicit, *not* necessarily justified, *theoretical* claim about introspection, which has managed to work itself into the framework of all strong phenomenal realist theories.

With such a wide range of intuitions about introspection, and with an apparent tendency to interpret what we find, when we look inwards, in the light of our (perhaps implicit) theoretical assumptions, it is far from clear whether we are on safe ground, if we make *any* proclamations about what it is that we know when we introspect the features of our sensory states, including qualia.

---

[47] And, as we have seen, the view also builds in the claim (which again must be introspectively based, if true) that this non-deducibility is so in a significantly stronger sense than the agreed, but far less surprising, sense in which it is so for indexical facts (note 36).

On the other hand, if we make no proclamations here at all, then we have no way of specifying our target of explanation as we try to understand qualia. Is there a middle ground? Is there a way to say *anything*, whilst remaining neutral as between competing theories of introspection? In the final section, I will argue that there is.

2.2.7 Some Moderate Subjective Properties

For my part, I am much more certain that there is *something* subjective about my mental life, and that I know this 'something' by introspection, than I am that what I know in this way transcends all physical and functional truths. Therefore, I am proposing that we allow ourselves to be guided, in our quest for qualia, by looking for an independently plausible account of introspection; specifically, we should look for qualia amongst the properties which are introspectible on such an independently plausible account[48].

I have just said that qualia are 'subjective' properties, but of course anything introspectible is subjective in a certain sense, for introspection consists in the ability of a subject to come to know properties of itself in a fundamentally first-person way (c.f. Section 3.6).

However, I am prepared to concede that some 'subjective' properties, in this sense, are the wrong type of thing to be qualia. Imagine, for instance, a subject seeing a red ball as a red ball (where red, in this case, should be thought of as a public, if gerrymandered[49], property). Essentially any account of introspection must allow that the right kind of subject can introspectively know *that* she is seeing a red ball when she is. This is a specific example of a general type of introspection, whereby a subject becomes aware that they have some 'propositional attitude'-type relationship (believing *x*, desiring *x*, seeing *x*, remembering *x*, imagining *x*, etc.) to some (perhaps only counterfactually existent) *public* object(s) or state of affairs *x*. I will be at least this much of a phenomenal realist: if independently plausible theories of introspection *only* allow that we have introspective knowledge of this type, then such theories do not have

---

[48] This does not amount to the requirement that qualia should always be introspectible. Whether or not non-introspectible qualia exist will hinge on the details of our theory of introspection, and on the details of any plausible candidate-properties for qualia within such a theory. For instance, on Shoemaker's account of introspection, mental states whose nature is to be introspectible can nevertheless exist in creatures which lack the resources to introspect them (Shoemaker, 1988 Section 3).

[49] This is Dennett's usage, it means that the outlines of what is and isn't red may depend on the constitution and interests of creatures like us, rather than on anything more fundamental about the world.

the materials to naturalise qualia. If things were to turn out thus, I should (and I think would!) accept that there *are* no qualia, and that I am as much in need of Dennettian therapy (Dennett, 1988) as are all those who maintain that qualia have non-naturalisable properties in the ways discussed in the earlier parts of Section 2.1.

But there seems a very natural next step to take, which is to wonder whether there might not be introspectible properties which are subjective in a slightly stronger sense: to wit, introspectible properties which cannot be fully specified, simply by specifying any number of the non-controversially introspectible properties just mentioned.

So now, imagine two subjects each seeing a red ball *as* a red ball. Imagine, also, that both have agreed on a common language for referring to public properties (red, ball, etc.) and to the 'propositional attitude' type states (including seeing *x*, etc.). Evidently things could be thus, even whilst there are facts about each subject's relation to the world which differ on a perfectly naturalistic account; for example, affective or motivational facts, and facts about the learnt associations between properties (e.g. red reminds one agent of blood and pain, and the other of celebration and good fortune). Now, these facts are subjective in yet a third sense: they are partly constitutive of the subject's relationship to the world. But what is not yet clear (at least, until we have an independently motivated account of introspection) is whether any such further facts can be known (perhaps, opaquely) in introspection. If they can be, then they are subjective facts in all three senses: subjective *qua* partially constitutive of the subject; subjective *qua* introspectible; and subjective in the sense just defined, of going beyond the most non-controversially introspectible facts.

In stating that the above is possible, I have not endorsed behaviourally undetectable inverted spectra: for the differences I have mentioned would all be behaviourally detectable. Even so, the situation described is not entirely unlike the standard inverted spectrum starting point. There could indeed be two subjects who see a red ball as a red ball (who even agree, in a shared language, that it is a red ball, and that each is seeing it) whilst there are *bona fide* introspectible facts about their experience which differ. As such, this seems to me a moderate approach with the potential to explain, rather than completely explain away, the widely held belief that qualia are invertible.

The suggestion that we concentrate on 'motivational, associative and affective' facts is just one proposal, intended to be compatible with the idea of being guided by an independently plausible theory of introspection. But there is a general problem with any proposal of this type, directly related to the two ways of understanding *a priori* analysis

noted earlier (Section 2.2.4.1). It could be taken to imply that the properties in question have been thoroughly "operationalized": that is, expressed in *fully* non-mental terms (setting aside the issue of whether or not this is truly possible). I have already suggested that that approach to *a priori* analysis leads to an overly strong reductionism which should be resisted. Indeed, if qualia are truly mental-level facts, then there is no reason to expect that anything which we know introspectively about them need entail any fully non-mental facts, *even if qualia can be explained on the normal scientific model* (remember that the water facts do not entail the H_2O facts). So the "operationalized" proposal is not the kind of proposal I am making.

Instead, the associative, motivational and affective facts (or whichever facts turn out to best fill the required role) should be read as properties at the independent mental level of description. The question at issue, when the proposal is read this way, is whether there is a conceptual independence between one type of mental description (a thinking, introspecting agent in a certain motivational state, say) and another (an agent having introspectible qualia, say). My suggestion is that we may well be able to find a two-way conceptual interrelation between qualia and the right set of not-so-obviously-qualitative *mental* concepts. If there is, then we would have a coherent account of the entire mental level, including introspectible qualia; and this whole account might yet map onto *some*[50] appropriate description of the physical in the normal way.

Of course, a standard response here is to claim that it is quite conceivable that our qualia are independent of any such (motivational, associative, affective, etc.) facts. Perhaps so, but I am not sure how (or indeed whether) I know that. I have suggested that the prior 'knowledge' of this 'fact', which many presume themselves to have, may be grounded in (implicit) endorsement of perhaps mistaken theories of introspection.

The strategy proposed here may also offer the possibility of explaining, rather than explaining away, other intuitions about the nature of qualia. I am thinking here, particularly, of Shoemaker's defence of a "moderate Cartesianism" (Shoemaker, 1988), which looks to be an entirely naturalisable account of a rather direct type of acquaintance we should *expect* to have with *any* introspectible property, on at least one independently plausible, apparently naturalisable, account of introspection.

---

[50] Lest I be misunderstood, I explicitly want to leave open the possibility that the currently popular information processing and representational descriptions may not be best suited for the low-level role in such an explanation.

Of course I need, for consistency's sake, to allow that my own presuppositions can be overruled. For any given prior intuition about the nature of qualia, if there are no facts which explain why this intuition was broadly (or even roughly) correct, then qualia do not have the intuited property. And, as I have already conceded, if none of our intuitions about qualia could be naturalised (not even the intuition that there *are* introspectible subjective properties, in the above sense), then there would be no qualia. But there does not yet seem to be any good reason to rule out the suggestion that we may find such properties, within some independently plausible account of the mental level in general, and of introspection in particular.

2.2.8 Summary

I have argued that what Chalmers calls *phenomenal realism* (Chalmers, 2003a) (and what I have called *strong phenomenal realism*) automatically rules out a certain standard form of scientific explanation. I have agreed with Chalmers that the modern phenomenal concept strategy cannot prevent this conclusion. Therefore, if Chalmers is right that the only way to "take consciousness seriously" (Chalmers, 1996) is to be a strong phenomenal realist, then a physicalist account of consciousness cannot succeed. This is certainly the case if physicalism is conceived of as a quest for this type of explanation of the nature of qualia, as I think it should be. But I have also briefly given reason to agree with Chalmers that physicalism cannot succeed on *any* reasonable interpretation, given these starting points.

I have then tried to throw doubt on the strong phenomenal realist starting point which leads to these objectionable conclusions. I have argued that whatever we know about the problematic aspects of qualia, which is supposed to lead us to strong phenomenal realism, must be known through introspection. I have noted that there is much evidence that we are entirely unclear about what we can introspect. I have also suggested that, historically, many theories of *perception* have built into themselves unjustified theoretical commitments as to the nature of *introspection*. I have argued that strong phenomenal realism (an account of the nature of conscious perception) may well be guilty of this same sin.

I have therefore proposed that we take a different approach, and have suggested that, as theorists, we should look for qualia amongst the properties introspectible on some independently plausible theory of introspection. I have noted that on essentially any theory of introspection, we can introspect certain 'propositional attitude'-style states,

including "seeing *x*" and "experiencing *x*", where *x* is some (at least counterfactually) public state of affairs. I have therefore defined 'subjective' properties, as those introspectible properties (if any) which can still vary (within or between agents), however many of the basic, uncontroversially introspectible propositional attitude style properties have been fixed. It follows directly from this definition that if there *are* such properties, they are *ipso facto* the right kind of thing to explain, rather than explain away, the inverted spectrum intuition. Not, that is, to explain the classic inverted spectrum, which remains incompatible with physicalism, but to explain how something which sounds very much like it is physically quite possible. I have also noted that such properties may be able to explain, rather than explain away, other apparently problematic intuitions about our epistemic relationship to qualia.

If we can find introspectible properties which are subjective in the above, moderate, sense, then we would have achieved some kind of phenomenal realism: there would be introspectible facts which at least come free of the standard propositional attitude facts about an agent. For the reasons given, it strikes me that such properties, if they exist, *are* plausible and adequate naturalizers of qualia. This is clearly not phenomenal realism as Chalmers defines it, but it does seem reasonable to call the present approach moderate phenomenal realism.

In sum, my proposal is that it is plausible and workable to *define* qualia as subjective, introspectible properties in the above moderate sense. Adopting this proposal allows us to be guided, in our attempt to understand qualia, by whatever independently plausible accounts of introspection we have[51].

However, at this stage we are certainly still entitled to ask whether or not there are any properties introspectible on Shoemaker's (or any other) model of introspection, and

[51] Might there be some yet more theoretically neutral definition of qualia? For instance, an analysis on which 'qualia' are whatever properties caused us to say that we had qualia in the first place (Sloman and Chrisley, 2003; Chrisley, 2008; Chrisley, 2009). I would claim that, if we found anything which matches some such more neutral definition, whilst not matching the stronger definition here, we would still say that there are no qualia (an example of an account in this territory is Dennett's broadly fictionalist analysis of consciousness: Dennett, 1991). Certainly, I think it is fair to say that the present definition captures *an* interesting aspect of the elusive concept 'qualia'; that it is a philosophically interesting question as to *whether or not there is anything which matches the definition offered here*, since if there is, it would naturalise central intuitions about qualia (in a moderate sense), and if there isn't, it would seem that such intuitions cannot be naturalised in any sense at all.

subjective in the sense outlined above. Even if there are, can such properties really explain our intuitions about qualia? All of this will be discussed over the course of the rest of the thesis. I will address specific claims to the effect that Shoemaker's account of introspection is independently *not* plausible in Section 3.6. Finally, I should like to note that the approach to naturalising qualia which I have just outlined here, and which I develop in more detail in Chapters 5 and 6, remains fundamentally different from Shoemaker's present approach. In Chapter 4 I will explain Shoemaker's own model, and explain why I think it cannot work.

2.3 Mind as Space of Reasons

2.3.1 Brief Introduction to the Notion

This thesis will concentrate on the notion of mind as physical locus of action for reasons (Sellars, 1956; McDowell, 1994; Hurley, 2003). A physical agent has a mind, in this sense, to the extent that the agent can be said to be acting (or, at least, able to act) for reasons. Throughout, the notion of action in question is an 'at least counterfactual' notion. That is, the actions in question are actions an agent either does take, or would take if only certain counterfactual conditions (not determining which actions are in question), obtained[52].

The discussion will start by using a fairly broad, intuitive notion of what it is to act for reasons, but this will be fleshed out in more detail in Chapters 3, 4 and 5.

The central concern of the thesis will be to defend the claim that mind, understood as action in a space of reasons[53], is all there is to mind: both as we know it when we observe others, and as we introspect it in ourselves. To defend this as it relates to *all* aspects of mind (affect, free will, qualia, 'mental representation', original intentionality, etc., etc.) would be far too ambitious a project. The specific aim here will be to defend

---

[52] This at-least-counterfactual formulation means that the action-based notion of mind can be applied to locked-in patients (Bauby, 1997; Laureys, 2005), for instance.

[53] In talking of *a* space of reasons, I do not mean to call into question the unitary notion of '*The* Space of Reasons' (c.f. McDowell, 1994). How not? Briefly, I do not believe that any of us, acting in *the* space of reasons, can have any reason to describe *another* agent as acting for reasons, except to the extent that we can situate its actions in the very same space of reasons in which we act. Because of this, I would argue that the notion of fully non-overlapping 'spaces of reasons' is not coherent (c.f. Davidson, 1974). Thus, when I talk of *a* space of reasons, I am indicating that subpart of *the* space of reasons to which an agent is sensitive. My thanks to Tom Beament for forcing me to clarify my usage of 'space of reasons' here.

the claim that this notion of mind alone is sufficient to naturalise qualitative, subjective, phenomenal feel.

As such, the claim that mind is to be understood as action within a space of reasons will be treated as a premise of this work, rather than as a conclusion to be defended. If a naturalisation of qualia based on this characterisation of mind (such as that attempted in this thesis) is doomed to fail, then either phenomenal feels must be eliminated, or we must accept that this notion of mind is far from exhaustive in its ability to match our (reasonable) pre-theoretic understanding of the mind. On this other hand, if the arguments in this thesis succeed, this can be taken as an implicit defence of this characterisation of mind.

2.3.2 Some Initial Objections to this Characterisation of Mind

Certainly some central cases of mental states are constitutively related to rationality, in a way which would have to be so if this notion of mind were exhaustive. Belief and desire are paradigmatically understood to be defined by their location within a realm of rational behaviour (Dennett, 1987). However, there would appear to be several *prima facie* plausible reasons for claiming that 'mind as locus of practical rationality' is in no way an exhaustive characterisation of the mental.

2.3.2.1 Rationality and Affect

For instance, in contrast to 'desire' understood in the above, perhaps somewhat technical, sense (as a state which is part of the *rationalisation* of certain actions), it is much less clear whether affect, more generally, can be understood as an aspect of rationality. Affect, or emotion, is a part of our everyday mental lives, and yet emotional behaviour is often paradigmatically contrasted with rational behaviour.

However, I believe that this is a mistake. Nothing can be identified as a space of reasons[54], unless it is a space of reasons for *action*. This, I will argue later (Section 5.2), means that no 'space of reasons' account of mind can be complete without ineliminable mention of affect. Indeed, such claims will be essential to the analysis of qualia to be presented in Chapter 5.

[54] Here (and frequently throughout), I use 'space of reasons' metonymously, to mean 'physical locus of rational action within some sub-part of the space of reasons' (see also footnote 53).

2.3.2.2 Imperfect Rationality

It might be thought that rationality is a poor model of mind, because so many aspects of real minds are clearly irrational. Very briefly, my response to this objection is that it is parallel to the flawed objection that the nature of belief and desire cannot be characterised in terms of their role in rational response, because so many real agents act irrationally given their beliefs and desires. In the case of this latter objection, the standard move is to argue that irrationality can only be identified as such, given a broad (and often underestimated) surrounding 'field' of rationality (Davidson, 1974; Dennett, 1987). This is widely, and I believe correctly, considered to defuse this objection.

It will be the burden of later chapters to argue that qualitative feel can be successfully analysed within a space of reasons account of the mental. But this no more implies that a creature with qualitative feels is perfectly rational, than the standard analysis of propositional attitudes implies that a creature with beliefs and desires is perfectly rational.

I do accept that I am nevertheless emphasizing rationality (albeit practical, embodied rationality) at precisely that point where it is often considered least relevant to our mental lives, as I will now point out.

2.3.2.3 Extra-Rational Sensation

It is widely thought that there is some viable notion of mere phenomenal sensation (conceived of as having no particular objective import in and of itself), which is far from being entirely characterisable in terms of its role in an agent's rationality (c.f. Smith, 2002). However, the notion of mind as locus of practical rationality cannot make space for such 'mere sensations'. The best this thesis can offer is a naturalisation of qualia as essential aspects of states whose nature is to present the world as being (or seeming to be) certain ways. For reasons outlined is Section 2.2.6, and taken up again at stages throughout, it is to be hoped that any felt need to allow for something yet more internal, subjective, and unrelated to the external world will be seen to amount to an at least implicit endorsement of unjustified *theoretical* claims about introspection.

2.3.2.4 Further Objections

If there are further objections to the claim that the space of reasons characterisation of mind is exhaustive, perhaps originating in other competing characterisations of the mental (free will, original intentionality, etc.), these will not be addressed here. I should perhaps clarify, in passing, that I suspect that quite the contrary is the case, and that a

space of reasons analysis of the mental is very well suited as a means of integrating the various apparently disparate ways which we have of characterising mind. However, space and time preclude any attempt to say more, here.

2.3.3 Experience as an Aspect of Practical Rationality

This thesis is centrally concerned with perceptual experience, and yet it may be considered unclear whether perceptual experience itself has any specific relation to rationality. It might be thought that I can 'have an experience', in *some* important sense, necessarily applicable when I consciously see something, without what is going on, in that sense, being a constitutively rational state[55]. Indeed, the most obvious version of such an objection would involve making the claim that I always have qualia, or mere sensations, when seeing, in combination with the common claim (Section 2.3.2.3) that qualia or mere sensations cannot be fully defined by their role in rationality.

In response to this objection, it should be noted that there *is* a conception of perceptual experience available, purely within the notion of mind as locus of practical rationality. According to this notion of perceptual experience, experience occurs when and only when public, worldly objects are present to the perceiving subject as at least potential reasons for action[56]. For instance, I am perceiving a tree, in this sense, when and only when an actual, worldly tree is present to me as an at least potential reason for action. It is this notion of perceptual experience which I will be treating as the central case.

This is obviously a notion of *veridical* experience: it can only apply when I am actually experiencing a publicly accessible object. However, it is often claimed, there is a perfectly valid *success-neutral* notion of experience: experience understood as that which is in common between perceiving something, and only seeming to perceive the same thing. This claim can be accepted, consistent with the approach developed in the

---

[55] The issues addressed in this subsection relate to a much wider debate about direct realism, disjunctivism and the nature of perception which is discussed a little further in Section 5.5.

[56] The object also has to be present in a certain way. Which way? The way characterised as characteristically visual by O'Regan and Noë (2001). However, the mere presence of 'mastery' of these sensorimotor contingencies is not enough – the 'mastery' has to be an integrated, partially constitutive part of a space of practical rationality before we have a 'presentation' of the object to a mind, as such. See the Appendix for related discussion.

present thesis, as long as the claim is read in a certain way[57]. As long as one sticks to thinking of mind as physical locus of action in a space of reasons (i.e. restricts oneself to only actual and counterfactual rational *behaviour*), it is quite possible to come up with a workable notion of success-neutral experience. To wit, a creature is having a success-neutral experience, whenever its actions *either* a) are veridically sensitive to an object in the world, in the way just sketched, *or* b) are *as if* the creature were veridically sensitive to such an object, when there is, in fact, some mismatch between any actual objects present and the creature's actions[58].

The above is, at least, a *usable* notion of success-neutral experience. Consider the case from the outside, looking at an experiencing subject. We could coherently choose to use the above definition to decide whether or not a creature were 'having an experience' (in the relevant, success-neutral, sense). It *would* be having such an experience whenever it was responding to things in the world, as for reasons, when and because the things it responded to were present in the right way. But it would also be having such an experience whenever it was acting *as if* actual, objective, publicly present properties and things were thus present to it, as reasons for action, when they were not[59].

There is nothing internally inconsistent with applying this notion: with defining success-neutral experience thus. But there would certainly be those who would say that some essential aspect of experience has been left out, in the choice to work only with these two *behaviourally* defined notions of it (the veridical and the success-neutral).

---

[57] In a way which doesn't, in fact, give such 'highest common factor' objectors what they want, see Section 5.5 for further discussion.

[58] What is in common is the same pattern of action in both cases. But *which* pattern of action this is cannot be defined in any way which is *more fundamental than* the definition of the pattern of behaviour involved in veridical experience. Hence, allowing that there is a perfectly valid success-neutral usage of *experience* in *this* sense is fully compatible with Hinton's disjunctivism (Hinton, 1973). Many disjunctivists have felt the need to claim that there is *no* valid success-neutral notion of experience – I believe that this is a mistake.

[59] This claim assumes that there is a certain way of identifying behaviour which is 'as if things were *visually* present', even when they are not (or, more generally, perceptually present). I suggest that the Noë/O'Regan sensorimotor contingencies (note 56) deal with this: something is visually present when these contingencies are satisfied, or when the subject 'takes it that they are' (which is to say, acts as if they are, or at least counterfactually would act as if they were).

Whether or not anything has really been left out will be under discussion. This section has aimed only to indicate which notion of perceptual experience one is limited to, if one limits oneself to treating mind as action within a space of reasons.

3. *The Nature of Introspection*

3.1 Introduction

The aim of this chapter is to present an analysis of the nature of introspection, and to argue that we do not introspect by means of coming to know any intrinsic properties of our minds. Equally (a slightly different claim) it will be argued that, having introspected, we cannot use the knowledge thereby gained to come to know any intrinsic properties of our minds.

More detail on what intrinsic properties are (in this context) will be given below (Section 3.2). But, briefly, acquiring intrinsic knowledge would involve acquiring knowledge which determines how things are, more specifically than would 'mere' knowledge of the mental relations between the introspecting subject and his or her own world. For instance, knowledge which determines (however opaquely or indirectly) that the introspecting subject is one particular internal physical state rather than another[60] is intrinsic knowledge.

This is contrary to a fairly common view on which introspection involves a process something like inner directed perception (again, more detail below). On such a view, introspection is achieved precisely by coming to know internal facts about one's state, which determine how things are with one more specifically than any mental-level relational characterisation can.

The denial of such views will be important for the subsequent discussion of qualia (Chapter 5). For if qualia can be introspected, and introspected properties cannot be intrinsic, then qualia cannot be intrinsic properties.

In order to argue for these results, an account of the nature of introspection, as a transition within a space of reasons, will be presented. This account has been argued for, relatively tersely, by Sellars (1956 Sections XII-XVI) and in much more detail by Shoemaker (see the papers collected in Shoemaker, 1996). On this account, the basic introspective transition, from having a mental state to knowing that one has it, is a

[60] As clarified below (Section 3.2), the mere fact that a subject is in a given, relational, mental state puts *some* constraints on the physical constitution of the subject; the claim here is that no knowledge *more specific than this* can be gained in or by introspection.

transition which certain agents simply can make: that is, such agents do not make the transition *by* doing something else (for instance, not *by* becoming aware of some intrinsic state). As will be clarified below, such an account is fundamentally inconsistent with the claim that introspection either involves, or consists in, any quasi-perceptual act.

Having presented Shoemaker's (Section 3.3) and Sellars' (Section 3.4) versions of the account, I will present a *prima facie* disagreement between Sellars and Shoemaker (Section 3.5), concerning what otherwise looks like a shared account. I will argue that there exists a resolution of this apparent disagreement. Showing how to resolve it helps to clarify that, although this account of introspection requires that the most basic acts of introspection be simple (that is, not further analysable) at the mental level, it certainly does not require that the physical mechanisms which enable such acts in a given agent need be simple (in the same, or any other, sense).

It will be noted that Shoemaker presents his work on introspection as a series of arguments *against* the view that introspection is quasi-perceptual, rather than as a positive account of some alternative view of introspection. Moreover, Shoemaker does not so much argue that quasi-perceptual introspection is *impossible*, as that it is *unnecessary*, given mere rationality in the self-ascription of mental concepts. This allows for responses such as those by Kind (2003) and Gertler (2003/2008), who argue that, although the form of introspection which Shoemaker discusses is possible, it is an over-intellectual form of introspection, and not the basic kind which occurs in us.

In Section 3.6, it will be argued that Shoemaker's position should be strengthened in response to these objections. I will argue that the correct conclusion which should be drawn from Shoemaker's (and Sellars') accounts is that quasi-perceptual self-knowledge, even if possible, *is not introspection*. I will also argue that the objection which states that the Shoemaker-Sellars account of introspection over-intellectualizes the process can be read in two ways, one of which rests on a misunderstanding of the account, and the other of which can be shown to be false.

In arguing against quasi-perceptual introspection, Shoemaker presents many separate arguments against the possibility of self-blindness: against the possibility, that is, of an agent who is rational in the self-ascription of mental concepts, but unable to come to know what an alleged quasi-perceptual mechanism of introspection is supposed to enable us to know. As a final step in the presentation of this chapter (Section 3.7), it will be noted that all these arguments share a similar form. It will be argued that this

similarity of form is no coincidence, and that there exists a generalised argument which leads to the conclusion that *any* non-intrinsic aspect of a space of reasons as such, is the right type of property to be introspected[61].

This chapter includes fairly extensive presentations of pre-existing work on introspection, but also several novel contributions. The presentation of existing work is considered necessary, because the claim that introspection is in no way perceptual is one which may well be counterintuitive to many readers who have not already been exposed to it. As such it may be useful to present the arguments for this claim in enough detail to be convincing.

Nevertheless, the latter sections of this chapter include several novel contributions, as follows: a novel resolution of an apparent tension between Shoemaker and Sellars (which assists in clarification of what is needed for introspection at the subpersonal level, on this account); novel responses to recent arguments against Shoemaker, by Kind and Gertler; the claim, which Shoemaker never clearly makes, that his arguments amount to a positive account of the nature of introspection, and not merely a denial of the quasi-perceptual model; a strong argument to the effect that this positive account has much more claim to be considered as introspection than does the quasi-perceptual account against which it is pitted; finally, an additional, novel extension of Shoemaker's account, to the effect that this analysis of introspection entails the introspectibility – by at least some possible agent – of *any* property of a space of reasons as such.

3.2 Intrinsic Properties

This brief section clarifies what is meant by 'intrinsic property', within the context of the present thesis.

The present work uses a view on which mind is seen as at least counterfactual action in a space of reasons (Section 2.3). On such a view, all mental states are relational states. Belief, desire, perception, even emotional states, are all described in terms of the (at least counterfactual) behavioural relation of the agent to aspects of its world[62].

---

[61] As will become clear, the sense of 'being of the right type to be introspected' being developed here *does not* entail that every creature which has states of this type has the ability to introspect them.

[62] It should be emphasized again that, on the version of this view being defended here, such 'behaviours' cannot be fully 'operationalized': cannot be fully re-expressed in equivalent, non-mental terms (c.f. Sections 2.2.4.1 and 2.2.7).

However, if we describe the behaviour of a physical agent at the action-for-reasons level, then there is still much we *haven't* said about the agent. For instance, we haven't said how much the agent weighs, or what colour it is, or, more generally, what it is made of. Of course, in saying that an agent perceives a tree in front of it (say), we have said *something* about the agent's physical structure, assuming a physicalist view of reality. For example, homogeneous matter (pure crystal diamond, say, or an unstructured gas) simply couldn't form an agent which behaves as for reasons, towards objects in its world, assuming anything vaguely like the laws of physics as we understand them. The point being made here is that, out of the vast number of physical structures of an agent which physically *could* explain a given pattern of behaviour in the world, merely ascribing some mental behaviour to the agent doesn't say *which* such structure is involved.

'Intrinsic' properties, as the term is used here, are simply those properties of an agent which are more specific than the relational mental properties. As such, the claim being defended in this chapter is that, when we introspect, we *only* discover such mental, relational properties. That, firstly, we do not discover our introspectible, relational mental states *by* discovering anything more specific about how we are constructed. And that secondly, having introspected, the knowledge we have gained (since only of relational mental properties) is not sufficient to determine anything intrinsic about our physical make-up.

3.2.1 Some Clarifications

There are a couple of *prima facie* objections to the validity of the usage of 'intrinsic' (as opposed to relational) just outlined, which should be dealt with swiftly before moving on to the discussion of the Shoemaker-Sellars model of introspection.

Firstly, if 'microfunctionalism' (Clark, 2001 p.36) were correct, if *no* detail about our physical-level properties truly cut below the mental level, then there would be no intrinsic properties, in the sense just defined. However, this is not a problem for the present thesis, which aims to argue *against* the view that such intrinsic properties play a role in characterising the mental life of an agent (i.e. against Churchland-Lewis type accounts of qualia as characterised in Section 2.2.4, and, as we will see, against Shoemaker's own current position on qualia as characterised in Chapter 4).

Secondly, many (including modern physicists) observe that modern physics treats *everything* as relational (Strawson, 1997 p.427; Smolin, 2000 pp.52/3). Nothing here

should be taken as contradicting that view. The 'intrinsic' properties discussed here are simply relational properties (to include properties such as 'being made of neurons', on a view on which this and *all* properties are relational) which are underspecified by the higher-level relational analysis which is the current topic of discussion.

3.3 Shoemaker's Arguments

3.3.1 Two Models of Perception

3.3.1.1 The Object Perceptual Model

Shoemaker's aim is to demonstrate that introspection is not like perception. In order to do this, he first needs to characterise perception. He offers two different (though related) characterisations: the *object perceptual* model and the *broad perceptual* model (Shoemaker, 1994a; Shoemaker, 1994b).

On the object perceptual model, one perceives facts by perceiving non-factual objects. For instance, one perceives the fact that the cup is on the table by perceiving the cup, and the table. To my own mind, this model of perception itself seems wrong, or at least very incomplete. How could perceiving the cup, and perceiving the table, explain the ability to perceive the fact that the cup is on the table? The ability to perceive the 'on' relation between the two certainly looks very like an additional ability, still entirely unexplained.

However it should be clarified that, on the object perceptual model of perception as Shoemaker presents it, the crucial point is that the ability to perceive such relations fundamentally *depends on* the ability to perceive the things thus related (Shoemaker, 1994a p.205). And perhaps this is quite a plausible characterisation of normal, everyday perception: we can only perceive the *on-ness* of the cup with respect to the table, by perceiving the cup and the table. That is to say, it is perhaps plausible (though still incomplete) to say that in normal everyday perception, seeing the *on-ness* requires the seeing of the cup and the table, indeed is partially constituted by the seeing of the cup and the table[63].

[63] Shoemaker recognizes that, for any given model of perception, there will be those who find that model of perception implausible. He proposes finessing this issue, if required, by treating his arguments as arguments against the claim that introspection conforms to certain common *stereotypes* of perception (Shoemaker, 1994a pp.203-204). Such arguments, he proposes, may be of interest even if perception itself does not in fact conform to those stereotypes.

Shoemaker presents several arguments to the effect that the entities accessible in introspection are not accessible as objects, in the same sense in which cups and tables are. He elucidates various aspects of the relation which one has[64], to the public objects of perception, and then presents a series of arguments, to the effect that we have no access to selves as objects in the same way (Shoemaker, 1994a Section III), nor to beliefs and desires (Shoemaker, 1994a Section V), nor to sensations or sense experiences (Shoemaker, 1994a Section VI).

The details of these arguments will not be presented here, partly because the present author has reservations as to whether the object perceptual model is a good model of perception[65], but mainly because they are not needed, in order to achieve the required results in this chapter. The results in question can be achieved simply by arguing that introspection does not conform to what Shoemaker calls the *broad perceptual* model (which will be presented next). This is so for two reasons.

Firstly, as Shoemaker himself lays out the positions, if introspection is not perceptual on the broad perceptual model, then it cannot be perceptual on the object perceptual model. This is because the object perceptual model is defined as the broad perceptual model plus additional conditions (Shoemaker, 1994a pp.205-208 & p.223). Arguments against the broad perceptual model are already arguments against the object perceptual model – as Shoemaker defines them both.

The second, and more important, reason for concentrating on Shoemaker's arguments against the broad perceptual model is that it is in these arguments that the basis of a positive analysis of introspection can be found.

3.3.1.2 The Broad Perceptual Model

On what Shoemaker calls the broad perceptual model (Shoemaker, 1994b), when one perceives, one comes to know something which has an existence independent of the existence of the means ('mechanism') whereby one comes to know it. As such, this

---

[64] Or is typically taken to have, see previous footnote.

[65] Nevertheless, the arguments Shoemaker presents seem interesting and important. Even if one has reservations as to whether perception of objects is exactly as Shoemaker characterises it, one can still feel that his lines of argument *would* still tell against the claim the selves (beliefs, desires, sensations) are knowable as objects, on whatever is the correct analysis of knowledge of public objects. Of course, formalising this intuition would involve developing this 'more correct' analysis of object knowledge, and showing how Shoemaker's arguments can be preserved within it.

broad perceptual model captures the standard intuition that the cups, trees, zebras, tigers and books of the perceptible world do not depend, for their existence, on the existence of the act of perceiving them[66].

Clearly, there are those who would think that this, too, is an incorrect model of perception. But in this instance it can be said (perhaps with more certainty than it can of the object perceptual model) that there is something right here, about this characterization of perception. Within the *mundane* framework within which there are cups and trees, and perceivers of them, we do not *normally* say that a given cup has no existence independent of the means whereby I perceive it. Even a metaphysical idealist need not take their idealism to falsify these claims of independent existence, *where these are only meant in the mundane sense*. If Shoemaker's arguments against the broad perceptual model are correct, then the targets of introspection[67] do not have an independent existence in any sense, not even in the mundane sense in which (even an idealist should concede) normal, public objects do.

In arguing against the applicability of the broad perceptual model to introspection, Shoemaker develops several arguments concerning the possibility, or otherwise, of what he calls 'self-blindness'. It is these arguments which will be presented in some detail in the remainder of Section 3.3, since they can be related quite directly to Sellars' position (Sections 3.4 and 3.5), and then used to develop a positive analysis of the nature of introspection (Sections 3.6 and 3.7).

3.3.2 Introduction to Self-Blindness

In arguing against the quasi-perceptual nature of introspection[68], Shoemaker asks us to consider the hypothetical case of an agent who is as rational as the rest of us in understanding the kinds of things which we come to know via introspection, but who is nevertheless incapable of introspecting.

Shoemaker calls such agents "self-blind", and he presents several arguments aimed at showing the impossibility of self-blindness. All his arguments have a similar structure:

---

[66] In fact, Shoemaker additionally characterizes the broad perceptual model in terms of the existence of *a causal mechanism which normally produces beliefs which are true*, about those things separate from it (Shoemaker, 1994a pp. 206 & 223). Once again, Shoemaker's aim is to deny that there is such a mechanism. Discussion of this part of the characterisation will be postponed until Section 3.5.

[67] That is to say, the thoughts, beliefs, experiences, etc. which become known in introspection.

[68] From here on, we will present arguments against the *broad perceptual* model, as just characterised.

they are designed to show that simply being rational in the appreciation of the knowledge which the quasi-perceptual mechanism of introspection is supposed to deliver, is itself enough to introspect, with no quasi-perceptual mechanism required. This, Shoemaker suggests, "calls into question" (Shoemaker, 1988 p.41) the supposition that the self-blind lack something which the rest of us possess; that is, it calls into question the notion that there is any quasi-perceptual mechanism of introspection in us.

The force of Shoemaker's arguments follows from two factors. Firstly, the states introspected (in all the cases which Shoemaker considers) are fully specified in terms of their role in an agent's rationality (i.e. they are constitutively rational states such as belief and desire). Secondly, to the extent that Shoemaker's arguments go through, such states can be introspected purely by the exercise of rationality. As such, the states introspected are not independent of the means of introspecting them, in the way required by the broad perceptual model of introspection: no separate mechanism of introspection is required.

In Section 3.6, we will see that Kind (2003) has objected that such arguments only show that we *need* not introspect via a quasi-perceptual mechanism, and not that we *do* not. In a not-unrelated vein, Gertler (2003/2008) has also objected that Shoemaker's arguments presuppose "an excessively high degree of rationality" in the introspecting agent.

The response to Kind offered below will be that Shoemaker's arguments do not so much 'call into doubt' the existence of quasi-perceptual mechanisms in us (though they do), but rather that they should be taken to show that quasi-perceptual self-knowledge, even if possible, *is not introspection*.

The picture of introspection being presented (and argued for) in this chapter will rightly seem implausible to many readers, as long as it seems as if quasi-perceptual self-knowledge would be mechanistically easier and cheaper for evolution to produce than the (allegedly) over-rational, 'replacement' self-knowledge which the self-blind must use.

As such, in the presentation of a sample of Shoemaker's arguments in this area, below (Sections 3.3.3 to 3.3.5), some emphasis will be placed on clarifying exactly how much rationality is required of an agent, in order that it have no need of a quasi-perceptual mechanism for self-knowledge. We will then return to the issue of the implications of the view for the physical mechanisms of introspection in Sections 3.5 and 3.6.

It might also be objected that Shoemaker's arguments cannot be extended to purely sensory or perceptual states since (it is often claimed) such states cannot be (fully) defined by their role in a creature's rationality. This important issue is mentioned in Sections 3.3.6 and 3.4.2.4, then returned to in more detail in Chapters 4 and 5.

3.3.3 Co-Operation With Another Agent

3.3.3.1 The Argument

Shoemaker asks us to consider a self-blind agent who wants to co-operate with other agents on some task (Shoemaker, 1988 p.40; Shoemaker, 1994b p.238). To see the kind of reasoning involved in Shoemaker's arguments, imagine that the agent believes that *P* is true, where *P* is relevant to the task. Imagine, also, that the agent has reason to believe that one of the agents with which it is co-operating does not believe that *P*. Now, "ceteris paribus, one is most likely to achieve one's ends if one acts on assumptions that are true" (Shoemaker, 1988 p.40). Since our agent's ends and the other agent's ends are the same, with respect to the task, then it will be in our agent's interest to let the other agent know that *P*; to say "*P*" for instance.

So far so good. But this is not enough to argue that an agent this rational cannot be self-blind. What we need to worry about is when, if ever, it will be rational for our agent to say, or think, "I believe that *P*".

So now we come to Shoemaker's version of the scenario, in which the agent in fact thinks that *P* (without any presupposition that it knows that it does so) and believes that the other agent also thinks that *P*, but has reason to believe that the other agent does not think that it (the first agent) believes that *P*. (It should be clarified that our agent *must* be able to *entertain* the thought that it believes that *P*, or it is not, as the hypothesis requires, as rational as we are; the agent is just supposed, for the purposes of *reductio ad absurdum*, not to be able to come to know, in a first-person way, that it believes that *P*.) Now, Shoemaker suggests, our agent:

> "could reason as follows. "*P* is true." [This expresses his belief, but it of course doesn't say that he has it.] It is therefore to anyone's advantage, by and large, to act on the assumption that *P* is true, for, ceteris paribus, one is most likely to achieve one's ends if one acts on assumptions that are true. Since this applies to anyone it applies to me – ceteris paribus it is to my advantage to act on this assumption. But that means acting as if I believe that it is true." (Shoemaker, 1988 p.40, with Shoemaker's own parenthetical insertion in the version of his paper referenced here)

This last sentence is the key move. Part of the *ex hypothesi* set up, concerning self-blind agents, is that they know what words such as 'believe' mean, in application to themselves. If self-blindness were a coherent possibility, then such an agent could understand the import of thoughts (and statements) containing such concepts applied to itself, but would nevertheless be at a loss as to whether or not such thoughts were true.

But here, Shoemaker aims to show that merely understanding the relevant thoughts, in application to itself, gives the agent premises sufficient to conclude, from *P*, that it is rational to act as if *P* (the first case), and as if it believes that *P* (the second case). This is enough to enable the agent to say "I believe that *P*" to the other agent, to use "I believe that *P*" in its own thoughts, etc., as and when this is true.

3.3.3.2 How Much Rationality is Required?

Shoemaker claims that the availability of such a line of reasoning throws into doubt the claim that we, in our own introspection, use some additional, quasi-perceptual, mechanism, above and beyond mere rationality, in order to know of our own beliefs. Now, it would be very easy to misunderstand Shoemaker as thereby equating introspection with *actually taking such a line of reasoning*. Nevertheless, this is *not* what Shoemaker means to require, of an introspector. A passage from another paper on the same topic makes this explicit:

> "The reason for pointing out that such reasoning is available is not to suggest that it regularly goes on in us – obviously it doesn't – but rather to point out that in order to explain the behaviour we take as showing that people have certain higher order beliefs, beliefs about their first order beliefs, we do not need to attribute to them anything beyond what is needed in order to give them first-order beliefs plus normal intelligence, rationality and conceptual capacity. What the availability of the reasoning shows is that the first-order states rationalize the behavior. And in supposing that a creature is rational, what one is supposing is that it is such that its being in certain states tends to result in effects, behavior or other internal states, that are rationalized by those states." (Shoemaker, 1994b p.239)

That is, just making the transition which is *rationalized by* the detailed line of thought is rationality enough. Still, we might wonder whether Shoemaker means to say that the explicit line of reasoning has to be at least *available* to (i.e. thinkable by) an agent which can introspect, just that it is not usually taken, in everyday introspection.

However, the quote above, especially in combination with other aspects of Shoemaker's writing, particularly on the applicability of his ideas to animal minds (Shoemaker, 1988 Section III) strongly supports a reading on which (for some creature,

of middling complexity) this explicit line of reasoning need not be thinkable at all, even whilst the creature can introspect. The creature needs no more (nor less) than the ability to simply make[69] the introspective transitions (from thought to meta-thought) which are *rationalized* by the relevant states.

3.3.4 Self-Knowledge and Desire

3.3.4.1 The Argument

Shoemaker also extends his arguments against the possibility of self-blindness to motivational states such as desire. Shoemaker suggests that if an agent (in this case, his self-blind man, 'George' – introduced for the sake of *reductio*):

> "… is capable of using language at all, he should be capable of giving linguistic expression to his desires, e.g., by making requests and other speech acts aimed at the attainment of things he wants. And if he is capable of doing this, he should be capable of learning to do it by saying things of the form "I want *X*" or "I would like *X*."" (Shoemaker, 1988 p.46)

This argument works as follows: a creature is *correct* to say, of itself, "I want *X*", when it does want *X*. Equally, where expressing a desire (e.g. making the noises "I want *X*" when appropriate) helps to attain the object of the desire (as it often will), then the creature is more likely to attain its desire if it can learn to make this public expression of desire, when applicable. As such, sufficient rationality alone is enough for an agent to be able to learn to make these 'noises', as and when they apply. But if these noises are being made correctly, as and when they apply, then they are not being used as mere noises, they are being used as words.

3.3.4.2 How Much Rationality is Required?

Once again, there is a line of reasoning which rationalizes the transition, from wanting *X* to asserting that it wants *X*, but the creature need not take it. The creature is already rational in its assertions of "I want *X*" if these assertions are ones which track its states of wanting *X*.

[69] Or simply learn to make – see Section 3.5 below.

3.3.5 Self-Knowledge and Moore's Paradox

3.3.5.1 The Argument

Closely related to Shoemaker's self-blindness arguments, are his many lines of argument linking Moore's paradox to self-knowledge (Shoemaker, 1988; Shoemaker, 1995). Moore's paradox concerns utterances such as "It is raining, but I do not believe that it is raining". There is nothing directly self-contradictory in such an utterance, for a speaker could utter it, and it could be true (as, for instance, if the first part of the conjunct is something which the speaker utters, but doesn't believe). Nevertheless, as Shoemaker says, in uttering such as sentence assertively "some sort of logical impropriety has been committed" (Shoemaker, 1988 p.34), "one could not hope to get one's audience to accept both conjuncts on one's say so, and could have little hope of getting them to accept either" (Shoemaker, 1988 p.35). The challenge, then, is to explain what the logical impropriety is in such statements, given that it is not outright self-contradiction.

Shoemaker points out that "it has been widely assumed that both the paradox and its resolution have to do with the linguistic expression of belief" (Shoemaker, 1995 p.74). He questions this:

> "What seems to me too little noticed is that there is something paradoxical or logically peculiar about the idea of someone's *believing* the propositional content of a Moore-paradoxical sentence, whether or not the person gives linguistic expression to this belief." (Shoemaker, 1995 pp.75-76)

Shoemaker presents several lines of attack, aimed at showing what is wrong with merely believing that which is expressed by a Moore paradoxical utterance. Rather than reconstruct these arguments in detail here, it will simply be noted that the basic line of argument is of the same form as that used in the previous two sections: there is always a line of reasoning which leads from believing something to believing that one does; therefore believing something *rationalizes* (in the sense discussed in Section 3.3.3.2) believing that one believes it. As such, it is not fully rational both to believe something, and to believe that one does not believe it. Thus Shoemaker's point follows: the logical impropriety is present merely in entertaining the thought which the Moore paradoxical utterance expresses.

3.3.5.2 How Much Rationality is Required?

Nevertheless, Shoemaker emphasises, the belief system of a real subject need not to be *fully* rational, *fully* self-consistent (Shoemaker, 1995 p.85)[70]. This need not undermine the argument above. For beliefs and desires are nevertheless *defined* by their role in a space of reasons. As has been often argued (e.g. Davidson, 1974; Dennett, 1987) error and failures of rationality can only be made sense of on the basis that beliefs are *typically* true, and the transitions between them *typically* rational. The line of argument above is not meant to demonstrate that real agents cannot be irrational – merely to show that believing the thought expressed by the Moore paradoxical utterance *is* irrational.

3.3.6 Self-Knowledge and Pain

3.3.6.1 Self-Blindness and Rational Response to Pain

Shoemaker does not say as much about the self-ascription of pain and other sensory states as he does about the self-ascription of more self-evidently rational states such as belief and desire. He concedes that he finds it "less obvious" (Shoemaker, 1990 p.71) how there could be a constitutive connection between rationality and introspective access to sensory states (especially sensations such as pain), as compared to the stereotypically rational propositional attitude states such as belief and desire, with which most of his arguments are concerned.

However, since later sections of this thesis (especially Chapter 5) will be centrally concerned with the relation between qualitative feel and rationality, it will be worth quoting what Shoemaker does say.

Shoemaker runs his self-blindness argument, for the case of pain. He asks us to:

> "try to imagine creatures who have intellectual, conceptual, etc. capacities comparable with ours, and who also have pain, but who are introspectively blind to their pains. Their only access to their pains is a third-person access – i.e., observing their own behavior, or their own inner physiology". (Shoemaker, 1994b p.227)

He emphasises a point here which is important for understanding his picture of introspection, and his arguments against the inner-perception picture:

[70] As Shoemaker makes clear in a footnote, the passages emphasizing this have been added to the revised version of the paper referred to here.

> "It must not be supposed that these creatures do not *feel* their pains. Pain is a feeling, and what they are self-blind to are, precisely, their feelings of pain." (Shoemaker, 1994b p.227)[71]

In discussing whether self-blindness to pain is possible, Shoemaker is discussing whether or not it is possible to feel pain, understand the concept of pain, have normal rationality, and be incapable of knowing (directly, non-inferentially) that one feels it.

Shoemaker asks us to consider whether the above supposition is really coherent. Can we make sense of pain being unpleasant, in such a creature? As Shoemaker points out, a normal consequence of such unpleasantness, is that the creature dislikes what is unpleasant, and wishes it would end. And this typically leads to behaviours (going to the medicine cabinet, phoning the doctor, etc.) (Shoemaker, 1994b p.227) *rationalized by* this wish for the pain to end. But if the creature takes these typical actions, and is to remain self-blind to its pain, it appears it must remain self-blind to the reasons for its actions. In such a case, it would seem to the creature "as if someone else had taken possession of its body" (Shoemaker, 1994b p.227). Once again, we can observe that it is perfectly possible for a real creature to be irrational. What would appear to be impossible is for a creature to be rational, and yet self-blind to its pain.

Elsewhere, Shoemaker says a little more about the link between pain and rationality:

> "Normally the behavioural effects of pain are partly a function of the subject's beliefs and desires. In leaping from the frying pan one tries to avoid leaping into the fire. The bodily protection system of which pain is a part exploits the rationality of the creature. Pain does not simply cause bodily movements apt to be advantageous to the creature [such as withdrawing one's hand reflexively from the fire]; it gives the creature a *reason* for acting in certain ways … . It is, I suggest, the fact that the explanatory role of pain is of this sort, i.e., that it is a reason giving role, that qualifies pain as a mental state. And I suggest that its playing this role requires that we have a special first-person access to it." (Shoemaker, 1990 p.71)

We might wish that Shoemaker had expanded on this theme (since what he does say appears to relate closely to arguments given in Section 5.4), but these comments occur within the context of the above-quoted admission that he finds it "less [than] obvious" how to fully extend his arguments to cover the case of sensory states. For this reason he says, "I shall limit myself to just one brief remark about this" (Shoemaker, 1990 p.71).

---

[71] Isn't Shoemaker begging the question, here? Can pains be felt without being perceived? It depends what you mean by pains (c.f. Section 5.4.4), but if the account of Chapter 5 is correct, pains *qua* feelings can indeed be *had* (felt; experienced painfully) without being perceived.

The brief remark consists in a single paragraph, about half of which has been quoted in the above.

The present thesis will later attempt to say considerably more, on the connection between felt, experienced pain and rationality.

3.3.6.2 Self-Blindness and the Unpleasantness of Pain

Shoemaker has argued that a state to which one is 'self-blind' to cannot lead to any of the normal behaviours rationalized by the presence of pain. But perhaps (he suggests, again for the sake of *reductio*) such a state could still be unpleasant. Could still have:

> "an intrinsic phenomenal character that constitutes its being unpleasant, and so makes it such that anyone who was introspectively aware of it would find it unpleasant" (Shoemaker, 1994b p.228)

But if this were possible, it would imply the further possibility that "all of what we take to be innocent pastimes produce in us states that are extremely unpleasant, but of which we are totally unaware" (Shoemaker, 1994b p.228), and equally the possibility of a subject whose "pains hurt, but they don't hurt *him*" (Shoemaker, 1994b p.228). Shoemaker conclusion is that any such proposal is too far from the normal meaning of these words to be made sense of (Shoemaker, 1994b p.228).

Of course – as Shoemaker also rightly concedes – there certainly could be states which play *part* of the role of pain, and to which we are self-blind, for instance:

> "[states] caused by … bodily damage of various sorts … and … causing behaviors, such as winces, grimaces, and moans, that can be involuntary and do not have to be seen as motivated or "rationalized" by beliefs and desires" (Shoemaker, 1994b pp.228-9)

But such a state, stripped of its link to any of the voluntary behaviour motivated by pain, "would not be pain. Indeed, it would not be a mental state at all" (Shoemaker, 1994b p.229). Or so Shoemaker suggests. This is a substantial claim. It can be partly supported by the arguments of Chapter 2, to the effect that phenomenal feels, of the type which we wish to explain when we talk about qualia, must be introspectible. It can be considerably further supported by the analysis of qualia which will be offered in Chapter 5. But I will briefly say a little more on this, now, in the next subsection.

3.3.6.3 Are Pains Really Rational States?

It may well be felt that such comments are very much in danger of over-rationalising mere felt, phenomenal pain. If so, the present discussion of the nature of introspection

can be treated, *pro tem*, as a discussion of the nature of introspection for more self-evidently rational states (beliefs, desires). Whether or not such an account can be extended to pains, and to phenomenal feels more generally, will then remain an open question until Chapter 5. I believe that the account of phenomenal feel given in that chapter justifies Shoemaker's brief remarks on pain, as quoted above. But the account given there is not Shoemaker's account. Indeed, it will be argued in Chapter 4 that Shoemaker's account of qualia is in unacceptable tension with his own account of introspection (which may perhaps explain why Shoemaker feels able to offer only the limited comments quoted above).

3.3.7 Summary of Shoemaker's View

Shoemaker never really says that he is offering a positive account of introspection. His aim is to show that introspection is not quasi-perceptual. He does this by showing that someone who is merely rational in the self-ascription of the mental concepts required to entertain introspective thoughts can already introspect.

The rational transitions which are sufficient for introspection, on this account, are not quasi-perceptual on the object perceptual model, because they do not involve the subject's accessing anything other than the mental state as such, in order to access the mental state (one does not introspect *by* doing, or seeing, or accessing, something else). Equally, these rational transitions are not quasi-perceptual on the broad perceptual model, because what is accessed (to wit, aspects of mind, construed as the realm of transitions in a space of reasons) is not independent of the mode of access (to wit, the exercise of mind, construed as the realm of transitions in a space of reasons).

However, if Shoemaker's arguments work, then he *has* offered a positive model: introspection consists in making such rational transitions. Or, at least, *one* form of introspection consists in making such transitions. We will return below (Section 3.6) to the question of whether our introspection (or, indeed, all introspection, truly said) is of this form.

3.3.8 Implications of This View for Knowledge of Intrinsic Properties

This account of introspection entails certain results concerning knowledge of intrinsic properties (Section 3.2). For what becomes known in introspection (on this model) are relational mental properties *as such*. We come to know properties like believing *x*, perceiving *x*, desiring *x*, etc., but we do not do so *by* coming to know (nor by being quasi-perceptually acquainted with) something else. As such, there is no room here for

any account on which we come to know about our mental, relational properties *by* coming to know about some other, intrinsic properties: it would be an example of the quasi-perceptual account which has been argued against.

Equally, since what is known in introspection (on Shoemaker's account) is purely relational, there is nothing about this knowledge which enables one to deduce *from* it the existence or nature of any more intrinsic properties of one's makeup.

As such, if this account of introspection is correct (which will be further defended in the remainder of this chapter, especially Section 3.6), then one cannot discover any intrinsic properties of one's makeup in or by introspection. More specifically, one cannot discover anything about the specific physical properties of one's makeup, nor can one discover any non-relational mental properties (such as qualia, *on the most common account of them*). This issue, and it's implications for an account of our phenomenal mental lives, is taken up in Chapters 4 and 5.

3.4 Sellars' Position

3.4.1 The Connections Between Shoemaker and Sellars

As we have just seen, Shoemaker has presented an extensive defence of the idea that introspection is unlike perception, and is instead a certain kind of noninferential transition within and about a space of reasons. The aim of the present section is to argue that the classic exposition of this very same view is to be found in the final sections of Sellars' *Empiricism and the Philosophy of Mind* (Sellars, 1956) (hereinafter 'EPM').

It might be doubted whether Shoemaker and Sellars have exactly the same view on this, since the form of words they each chose, to discuss a central aspect of the view, makes it sound as if they directly disagree with each other about that aspect. This disagreement will be presented below (Section 3.5), and it will be argued that the disagreement is only apparent. The explanation as to why will help to clarify the relation between the personal and subpersonal levels, within this analysis of introspection.

As well as the value gained in addressing this apparent conflict, it will also be helpful to present (if only in overview, rather than in exhaustive detail) another set of arguments for the perhaps rather counter-intuitive notion of introspection on which the rest of this thesis relies.

3.4.2 The Myth of Jones

3.4.2.1 Jones' Theory of Thought

Sellars' account is to be found in the later sections of EPM (Part XII, Section 48 onwards). In this extended passage, Sellars' hero, Jones (EPM §60) undertakes what Sellars elsewhere describes as a "momentous experiment" (Castañeda and Sellars, 1961-1962/2006). In this "piece of science fiction" (EPM §48) Sellars imagines a fictional time in the past, the time of "Our Rylean Ancestors" (the title of EPM Section XII, where this presentation begins). They are named 'Rylean' in somewhat ironic homage to Gilbert Ryle's philosophical behaviourism:

> "the philosophical situation it [the thought experiment] is designed to clarify is one in which we are not puzzled by how people acquire a language for referring to public properties of public objects, but are very puzzled indeed about how we learn to speak of inner episodes and immediate experiences" (EPM §48)

These "talking animals" (EPM §49) are supposed to be in a state in which they have mastered essentially all of normal language, except for concepts for mental states[72]. Thus, for them, there are no mental states – their realm of understanding does not yet include such things.

But these Rylean ancestors can already speak to each other. Indeed, they possess the concept of speech – thus one might say to another "last Wednesday, I said to you that I was going to meet you this Tuesday". Further, these Ryleans understand the normal connections between speech and action: that, for instance, when someone says they will meet someone else next Tuesday then, all other things being equal, they will meet that person next Tuesday. The Ryleans explicitly know all this: for instance, they would explicitly aver this if asked. As Sellars says:

> "Let it be granted, then, that these mythical ancestors of ours are able to characterize each other's verbal behavior in semantical terms; that, in other words, they not only can talk about each other's predictions as causes and effects, and as indicators (with greater or less reliability) of other verbal and nonverbal states of affairs, but can also say of these verbal productions that they mean thus and so, that they say that such and such, that they are true, false, etc." (EPM §49)

---

[72] Sellars supposes a very tight connection between concepts and language, a connection which can arguably be loosened, although I only say a little more about this (see the Appendix, note 162).

What the Ryleans don't have, at this stage, is the ability to think that they think anything. They only have the ability to think that they say and do things. That is, they have concepts of 'say' and 'do' (and 'mean'[73], 'speak', 'true', 'false', etc., as above), but they don't yet have concepts such as 'thought', 'experience', 'sensation'.

In Sellars' story, what happens is that the inspired Jones (EPM §53) develops a theoretical model to account for all this doing and saying. On Jones' model, doings and sayings are the culmination of inner processes – in particular of the process which we would call (and which Jones chooses to call) thought. Thus, thoughts are theoretical entities which Jones can use to help him explain others' behaviour; in particular to help him explain their "intelligent nonhabitual behaviour" (EPM §56), i.e. to help him explain the things they do which are not just reactions to stimuli – to explain those actions which we (and Jones) describe as involving 'thinking'.

At this point in his account, Sellars insists that Jones understands such 'thoughts' by analogy with speech. Thus, that Jones is thinking of thoughts as 'inner speech'. But then, at first sight confusingly, Sellars says:

> "It is essential to bear in mind that what Jones means by "inner speech" is not to be confused with verbal imagery. As a matter of fact, Jones, like his fellows, does not as yet even have the concept of an image." (EPM §56)

The point being made here is that at this stage in his theoretical development, Jones does not have a notion of experienced inner states, and thus he is not thinking about (could not be thinking about) mental states in this way. At this stage, the 'thoughts' in Jones' theory are strictly theoretical entities. They are modelled on speech, they are inner (in as much as covert), but they are *not* "inner speech" *in that sense* where this phrase means verbal imagery.

The analogy with speech in Jones' theory has already bought him a lot, however, for it "*carries over to these inner episodes the applicability of semantical categories*" (EPM §57, original emphasis). Thus thoughts can *mean* this or that; can be *about* this or that; the content of given thoughts can be *true* or *false*.

3.4.2.2 Jones and the Introspection of Thought

In section 59, Sellars begins his account of introspection as such:

[73] At this stage, for a sentence rather than for a subject.

> "[O]nce our fictitious ancestor, Jones, has developed the theory that overt verbal behavior is the expression of thoughts, and taught his compatriots to make use of the theory in interpreting each other's behavior, it is but a short step to the use of this language in self-description. Thus, when Tom, watching Dick, has behavioral evidence which warrants the use of the sentence (in the language of the theory) 'Dick is thinking "p"' (or 'Dick is thinking that p'), Dick, using the same behavioral evidence, can say, in the language of the theory, 'I am thinking "p"' (or 'I am thinking that p.') And it now turns out – need it have? – that Dick can be trained to give reasonably reliable self-descriptions, using the language of the theory, without having to observe his overt behavior. Jones brings this about, roughly by applauding utterances by Dick of 'I am thinking that p' when the behavioral evidence strongly supports the theoretical statement 'Dick is thinking that p'; and by frowning on utterances of 'I am thinking that p', when the evidence does not support this theoretical statement. Our ancestors begin to speak of the privileged access each of us has to his own thoughts. *What began as a language with a purely theoretical use has gained a reporting role*." (EPM §59, original emphasis)

This is a very compressed section, and exactly what Sellars intends here is made considerably clearer in Sellars' correspondence with Hector Castañeda about this paper (Castañeda and Sellars, 1961-1962/2006). It turns out that a lot hinges on Sellars' cryptic "need it have?", in the above. This will be discussed below (Section 3.5).

For now, though, note that Sellars assumes that Dick (standing in for any agent who can learn to introspect) is trainable such that he can indeed learn to self-ascribe the terms of Jones' theory, as and when they apply to him (not necessarily always and perfectly, but in the main, with a fair wind, etc.). In Section 3.5, we will return to a discussion of exactly what is required for such trainable introspective self-ascription to occur.

3.4.2.3 Jones and the Introspection of Looking and Seeing

We have just seen that Sellars proposes that 'thought' can be understood on the model of covert speech. For instance, the state of *thinking that Paris is the capital of France* is the state introduced as a theoretical entity to explain actions including (but not limited to) saying that Paris is the capital of France.

Our Rylean ancestors are equally supposed able to understand statements involving publicly verifiable relations between subjects and objects such as "*seeing that the table is brown, hearing that the piano is out of tune,* etc." (EPM §60, original emphasis). This is why Sellars feels able to say, with very little ado, "among the inner episodes which

belong to the framework of thoughts will be perceptions" (EPM §60): Jones' theory extends very naturally to postulate sometimes covert states of the subject, which could be labelled "seeing that the table is brown" introduced to explain the behaviour which these Ryleans already know about, which occurs when a subject (with eyes open, etc.) is confronted with a brown table. Such states will also be introspectible on the Sellarsian model of the introspection of thoughts as just outlined.

It should be noted that this entire extension of Jones' theory, to the case of being able to say, on an introspective basis, "I see that the table is brown", is covered in the first short paragraph of the final main section (XVI, which is §60-§63) of EPM (which is entitled: "The Logic of Private Episodes: Impressions").

Moreover, Sellars leaves it almost entirely implicit as to what is involved in understanding (and introspecting) "looks" statements ("it looks to me as if the table is brown"), although he certainly does (by the start of §62), take it as demonstrated that such statements are introspectible, on the same model as other thoughts[74], when he talks of "such introspectible inner episodes as *its looking to one as though there were a red and triangular physical object over there*" (EPM §62, original emphasis). However, I think we may infer from what Sellars does say, that we are supposed to understand the theoretical internal state defined as something's *looking* a certain way, to be modelled on the publicly observable relation wherein a subject reacts (in appropriate ways) *as if* they were seeing a brown table (say), but where this particular ascription ('looks') is neutral as to whether the subject is in fact seeing a brown table or not.

3.4.2.4 Jones and the Introspection of Sense Impressions

The above extension of Jones' account (to the case of introspective ascription of *seeing*, *looking as if*, and so on) is all that will be required to discuss the introspection of perceptual states, on the theory being developed in the present thesis. As such, it might seem more than a little confusing that Sellars' covers all this so briefly, within what is a considerably more extended part of his paper (i.e. XVI, §§60-63) devoted to discussing the logic of (and eventually the introspection of) sensory experience.

In turns out that the main work of the final part of EPM is to allow Jones to develop a *different* (though related) theory from that so far discussed. Not a theory of thought

[74] It is right to say 'other thoughts', here: I have only just quoted Sellars stating that "perceptions" belong to "the framework of thoughts". See the next sub-section for further clarification of what is going on at this point in EPM.

(conceived such as to *include* states such as *seeing that* and it's *looking that*), but a theory of "sense perception". Sellars emphasizes the difference in subject matter between these two theories:

> "It cannot be emphasized too much that although these theoretical discursive episodes or *thoughts* are introduced as *inner* episodes – which is merely to repeat that they are introduced as *theoretical* episodes – they are *not* introduced as *immediate experiences*. Let me remind the reader that Jones, like his Neo-Rylean contemporaries, does not as yet have this concept." (EMP XV, §58 point 5).

He emphasizes the same point in a different way later on, talking of:

> "the assimilation of impressions to thoughts, and thoughts to impressions which, as I have already pointed out, is responsible for many of the confusions of the classical account of both thoughts and impressions", (EPM §61 point 1).

As such, what we have in EPM XVI is the process of Jones' developing a *further* theory of 'immediate experience' as such. On this theory, sense perception is seen as consisting in:

> "[sense] impressions, … which are the end results of the impingement of physical objects and processes on various parts of the body" (EPM §60).

Now we may certainly suppose that Sellars is merely using Jones to talk about what Sellars himself thinks is required to understand the logic of, firstly, thoughts, and secondly, sense impressions. Indeed, Sellars says as much, stating that Jones' theory is intended to:

> "throw light on the logic of our ordinary language about immediate experiences" (EPM §60).

However, it should be made clear that, from the point of view of the present thesis, Jones (and hence, Sellars) *has gone wrong* in postulating something further – to wit, sense impressions – where these are to be taken as something understandable and introspectible on a *different* model from that involved in understanding and introspecting states of seeing, looking, etc. Space and time preclude proper elaboration of this claim, as it applies specifically to Sellars' work. Nevertheless, it is hoped that it will become clear over the course of the thesis why I object to this analysis of sense impressions. Equally, it is hoped that the above is enough to pin down Sellars to having said something which I do indeed disagree with (see also footnote 75).

None of this weakens my claim that Sellars and Shoemaker share an account of introspection since, from my own point of view, Shoemaker himself also goes wrong

when it comes to accounting for the characteristically sensory aspect of perceptual experience (as discussed in Chapter 4)[75].

In any event, what I mean to endorse from both authors is their shared account of the introspection of *rationally* characterised states; and *not* their approaches to building upon this to account for introspection of *sensory* states. I will argue that no such additional material is needed.

3.4.3 Methodological Behaviourism and Introspection

Whatever Sellars thought about *sense impressions*, I believe that what he said in regard to those mental states which he characterises as aspects of *thought* as such should be taken to apply to all mental states:

> "that the fact that overt behavior *is* evidence for these episodes *is built into the very logic of these concepts*" (EPM §59).

The claim that this applies to *all* mental states will be further defended below, in the context of the analysis of qualitative feel to be offered here (Chapter 5).

To end this presentation of Sellars' account of introspection, one final quote from Sellars will be given. This is for two reasons. Firstly, it is further evidence of great commonality between Sellars and Shoemaker on introspection. But secondly, it helps to clarify what is and isn't entailed by this oft-rejected notion, that our mental concepts are fully public:

> "If we permit ourselves to speak of this privileged access to our states of mind as "introspection," avoiding the implication that there is a "means" whereby we "see" what is going on "inside," as we see external circumstances by the eye, then we can say that Behaviorism, as I shall use the term, does not deny that there is such a thing as introspection, nor that it is, on some topics, at least, quite reliable. The essential point about 'introspection' from the standpoint of [methodological] Behaviorism is that *we introspect in terms of common sense mentalistic concepts*." (EPM §53).

To this might be added what is certainly at least implicit in the above: such common sense concepts are fully definable in terms of at least counterfactual behaviour[76].

---

[75] Sellars further states that 'sense impressions' are "intrinsic" (EPM §61, points 2 and 3) and, it would seem, representational (EPM §61, point 3); these points also bear comparison with Shoemaker's current account of qualia.

[76] The behaviour in question may not be reducible to non-mental behaviour (c.f. note 62). Sellars clearly sees the need for non-reductive definition of the terms of Jones' theory, see EPM §61 point 3 and §55.

Simply stating this will not do, of course. Subsequent chapters will defend it, both by fleshing out such an account, and by an attempt to show that qualia (mental properties which have often been supposed to outrun such analyses) can in fact be captured by such an analysis, and thus introspected on the above model of introspection as it applies to states of a space of reasons *as such* (i.e. without mention of any further, intrinsic properties).

3.5 Shoemaker vs. Sellars?

3.5.1 Introduction

There certainly seems to be a lot on which Shoemaker and Sellars agree: that the facts about oneself which become known by introspection are not perceived; neither are they discovered *by* perceiving some other facts or states of affairs; that introspection is a basic personal level transition, which certain subjects 'just can' make – not by doing anything else, as far as the personal level goes.

But if we read Sellars' clarification of his terse "need it have?" (in the passage quoted in 3.4.2.2 above), in Sellars' correspondence with Castañeda (Castañeda and Sellars, 1961-1962/2006) – and if we also look at certain things Shoemaker has said, in defence of his own position – we can find a pair of passages which look to be in direct conflict, as regards what is involved, at a subpersonal level, in making this personal-level transition.

3.5.2 Castañeda's Colony of Viruses

As outlined above, in the central portion of Sellars' Myth of Jones, Jones trains one of his fellow humans, Dick, in the self-ascriptive use of the theoretical terms of Jones' new theory. Sellars supposes that "Dick can be trained to give reasonably reliable self-descriptions, using the language of the theory, without having to observe his overt behavior" (EPM §59). In correspondence with Sellars, Castañeda objects to this account. His problem is that Sellars, far from giving an account of introspection as a *non*inferential transition, appears to be relying on Dick's ability to make rational *inferences* from his behaviour to his own mental states:

> "Compare the following case which is on all fours like your theoretical inner episodes: …
>
> "Dick shows all the signs (criteria, symptoms, what you care to call them) of a person with a colony of certain filterable viruses lodged in his left kidney. He is taught the theory of viruses

so that he can infer from his signs that he has a colony of the viruses in question. He is able, then, to make the *theoretical* statement "I have a colony of viruses…".

"Now, if that is what happens in the case of [thought], we would have to say that Dick is conditioned to utter "I am thinking that-p" or "I have just had a thought that-p" on inspection of his behavior and circumstances. But this is just what we do not want." (Castañeda to Sellars, April 13, 1961)

In the same passage Castañeda accepts that there is a sense in which, *once* the person is trained, they no longer make the explicit inference. The problem which Castañeda highlights is that the type of training just described can only occur when and because the subject is able to explicitly notice the public signs.

It is important to realize that Sellars' agrees that this is *not* the kind of 'introspection' which we want. The process Castañeda describes, which is fundamentally *inferential*, is often supposed to be all that is available, in the way of 'introspection', on a behaviourist theory (c.f. Kind's objection to Shoemaker below, 3.6.3). But it is *not* the process which Sellars' meant to endorse, in his myth of Dick and Jones.

Sellars responds to Castañeda by questioning Castañeda's assumption, that it is only possible to train a subject to report on some state of affairs φ if the subject can already observe adequate signs of φ.

"It turns out that for some states φ (but by no means for all) we can bring about a connection between being in state φ and saying "I am in state φ"" [even though the subject can] "[neither] observe that he is in state φ … [nor] observe that he is in a state which is a sign of state φ." (Sellars to Castañeda, November 14, 1961).

Sellars continues:

"Roughly the difference between the cases where it can be done and the cases where it can't is that in the favorable cases, being in state φ, causes neural impulses which feed into the central nervous system in such a way that they can be hooked up with the neural processes which culminate in the utterance of "I am in state φ", i.e. (on my view) with the thought **I am in state φ**. … The neural impulses … need not be accompanied by sensation or feeling." (Ibid.)

And further:

"If we modify your example … by supposing that virus colonies oscillate between growing rapidly and decreasing rapidly in number … then my point is that Dick, by analogy with my myth, is trained to say "I have a rapidly growing colony of viruses" *when such rapid growth occurs*, and not *when he notices observable signs of such growth*." (Ibid.)

I think we can therefore see that at lot of thought is packaged into the short phrase "need it have?", in Sellars' original paper. Sellars' myth requires that Dick is 'wired up' such that he *can* be trained to respond with the meta-thought (which might be expressed in words as) 'I am thinking T' when the thought T occurs. Indeed, it is precisely because Dick does not make this transition by observing any signs, that we call the process introspection. Although Dick has been trained to use a theoretical vocabulary which is intrinsically third-person (outwardly directed), he has been trained to use it in an *intrinsically first-person way*.

The problem with Sellars' response, for present purposes, is that in clarifying his position, Sellars has made it clear that Dick must use a special (trainable) form of access which he *need not have had* (if not, he would not have been able to introspect). This appears to make Sellars' account match badly with Shoemaker's, for Shoemaker wanted to argue for the impossibility of self-blindness: of someone being rational in their self-ascription of mental concepts, and yet lacking this special access required for introspection.

Sellars more than once explicitly denies that the form of access in question is "perceptual or quasi-perceptual" (EPM §47) (the same point is made in the quote from EPM §53 in Section 3.4.3 above). To that degree, he matches Shoemaker exactly. But still this access looks like something which might be absent in an otherwise rational agent; it looks as if self-blindness is a logical possibility, on Sellars' account.

3.5.3 Shoemaker's Blood Pressure

Shoemaker cannot allow any mechanism which might have been absent in an otherwise rational agent. This would run directly counter to his arguments against the possibility of self-blindness. Shoemaker is well aware that he cannot allow any such mechanism. Worse than this, from the point of view of the present attempt to argue that Shoemaker and Sellars have the same view on the nature of introspection, Shoemaker discusses this very issue in terms which sound like a direct repudiation of Sellars' points above[77]. Shoemaker says:

---

[77] No explicit reference to Sellars is given by Shoemaker, in the paper from which the following quote comes. Indeed, neither EPM as such, nor Sellars' Myth of Jones, is referenced in *any* of the papers collected in Shoemaker (Shoemaker, 1996). *Science, Perception and Reality* (Sellars, 1963) (in which EPM was reprinted) is referenced twice, but only as regards Sellars' distinction between the "scientific image" and the "manifest image".

> "[I]t seems perfectly conceivable that there should be creatures … who have a "special access" to physical states of themselves which is not … mediated by sensations and background beliefs. We can imagine, for example, that the blood pressure of these creatures varies from one moment to the next, but that if you ask one of them what his blood pressure is he is always able (after some preliminary training) to answer correctly, and is unable to give any account of how he is able to do this, except by saying that once the question is put to him he "just knows" … . The anti-Cartesian, as I am conceiving him, sees no important difference between the special access we in fact have to our own mental states and the access these creatures would have to their blood pressure … . And it is on his view just a contingent fact that we have one sort of access and not the other; logically speaking, it could just as well have been the other way around." (Shoemaker, 1988 pp.26-27)

This sounds like a direct repudiation of Sellars' position, for Shoemaker certainly means to oppose the 'anti-Cartesian' of this passage[78]; and the language Shoemaker uses, to describe the access which the anti-Cartesian supposes we have to our mental states, is pretty much exactly the language Sellars uses, to describe the access which Sellars thinks we actually do have.

3.5.4 Resolution

How, then, should we resolve this apparent conflict between Shoemaker and Sellars? Is Shoemaker right to say that introspection is nothing like trainable, noninferential access to inner states? Or is Sellars right to say that this is exactly what introspection is like?

In fact, it will be argued here, both claims are right, and they need not be read as contradicting each other. For learning to introspect *is* like learning to noninferentially detect an internal bodily state, in the respect which Sellars means to emphasize. But it is also crucially *unlike* noninferentially detecting an internal bodily state, in the respect which Shoemaker means to emphasize. The separate points with each author is making can stand together, for they are not contradictory.

To see how this can be so, take, first, the following quote from Shoemaker:

[78] His aim is to defend a limited form of Cartesianism: that is, to claim that it is of the nature of mental states that they be knowable by introspection. There will be some more on what this limited Cartesianism comes to, in the defence of the claim that we have a special access to qualia (though no more or less special than the access we have to our other mental properties) in Section 6.2.

> "here[79] the utility of self-knowledge depends crucially on its being acquired by self-acquaintance; if I had to figure out from my behaviour what my beliefs, goals, intentions, etc. are, then in most cases it would be more efficient for others to figure this out for themselves than to wait for me to figure it out and then tell them about it." (Shoemaker, 1988 p.28)

This passage demonstrates that Shoemaker, like Sellars, is *not* claiming that sufficient (and sufficiently fast) rationality based on third-person observation is equivalent to normal self-acquaintance (i.e. introspection). Rather, he is claiming that introspection is something else: a first person way of correctly applying (what are equally) third-person concepts.

What, though, is involved in making such an introspective transition? All of Shoemaker's arguments tend to the conclusion that learning to make such a transition is *no more nor less* than learning to be appropriately rational in self-ascription of the relevant, publicly applicable, concepts.

As such, Shoemaker is quite right to emphasize that there is an "important difference" between noninferential access to blood pressure (were we to have it) and noninferential access to thought. The difference is not that one involves an internal mechanism, and the other does not; nor is it that one involves training and the other does not. Instead, the difference is the following. The first case involves gaining knowledge of states logically independent of thought itself (there can be blood pressure without thought; and thought – of some possible agent – without blood pressure). Whereas the second case involves gaining knowledge of thought itself: it involves thought about thought, where the transition (from thought to meta-thought) is not mediated by thought about something else.

It is *only this latter point* (that the transition is not mediated by thought about something else) which Sellars meant to emphasize, in stating that noninferential self-knowledge of a virus colony as a good analogy for self-knowledge of the mental. On the other hand, Shoemaker sees disanalogy where Sellars sees analogy *only because Shoemaker is concentrating on something else*: the (logical) independence, or otherwise, of what becomes known, from the means of knowing it.

---

[79] The specific quote given concerns the utility of self-knowledge in the context of co-operative endeavour, but I think we can fairly take it that the comment is generally applicable on Shoemaker's view.

Thus, Shoemaker and Sellars can (and I believe would) both agree that self-knowledge *is* like the virus/blood pressure example in the sense in which Sellars meant the comparison (trainable, non-inferential self-knowledge, *not* based on third-person observation), *and* that it is not like this example, in the sense which Shoemaker emphasizes (self-knowledge of thought is essentially more intimate than self-knowledge of a non-mental state).

3.5.5 A Mechanism for Introspection?

The above discussion demonstrates that this shared view of both authors, on the nature of introspection, is entirely compatible with there being certain physical facts about a creature which explain its ability to introspect. Indeed, I am sure that both Shoemaker and Sellars would accept that physicalism entails that there *must* be such physical facts, and would both endorse physicalism (as would the present author).

Note, however, that the present discussion of introspection has not been about such mechanisms, it has been about getting clear as to what introspection is, at its own level of description: i.e. at the personal level.

Perhaps such explicit personal-level analysis is not a precondition for looking for the subpersonal mechanisms which enable introspection. On the other hand, it is to be hoped that such analysis is far from irrelevant: *ceteris paribus*, one's chances of identifying and understanding the mechanisms underlying *x* are greater, the greater one's understanding of the nature of *x*.

For a little more on the implications of this view for a mechanistic understanding of introspection, see also Section 3.6.4.3 below.

3.6 Why Shoemaker's Claim Should Be Strengthened

3.6.1 Introduction

The above amounts to an account of introspection shared by Shoemaker and Sellars. To be clear, this account is also endorsed (though with the reservations mentioned in 3.4.2.4) by the present author. It is fair to say that the account is only tersely presented by Sellars. Much more extensive arguments for this account are given by Shoemaker, and we will concentrate, for the rest of this chapter, on Shoemaker's presentation.

There are, though, dialectical rather than substantive problems with Shoemaker's approach: it is not that there is anything incorrect in his underlying position on introspection, but there are (I will argue) two 'weaknesses' in Shoemaker's presentation

of his position. These are weaknesses only in as much as that they are liable to lead to misunderstanding (and have indeed lead to it, as we will see). No fundamental falsehood or incorrectness is involved.

The first problem is an over-emphasis on the lines of thought which rationalize the introspective transitions which Shoemaker talks about, coupled with an under-emphasis on the fact that an agent need not be able to *follow* such lines of thought in order to simply make the rational transitions in question. As already emphasized above, 3.3.3.2, Shoemaker does make this point, but he makes it unemphatically and infrequently enough that it looks to have been a common cause of misunderstanding of the account.

The second, perhaps more serious, weakness in the presentation of his position, is Shoemaker's way of arguing as if purely *against* quasi-perceptual introspection, rather than *for* some more positive model.

Below, we will look at objections to Shoemaker's account by Gertler and Kind. Gertler's objection has two readings. On the first reading, it amounts to no more than a misunderstanding of the account related to the first issue mentioned above. On the second reading, it raises the same points as Kind's objection. Kind's objection (and Gertler's on the second reading of it) throws up a genuine weakness in Shoemaker's presentation, and require that Shoemaker's arguments be strengthened, to emphasize and defend the claim that he has given a positive model of introspection, not just a set of arguments against a quasi-perceptual model.

3.6.2 Gertler's Objection

3.6.2.1 The Reading Which Amounts to a Misunderstanding

In reviewing the rationality model of introspection (Shoemaker's model is one of two examples she gives), Gertler worries that:

> "proponents of the Rationality model … [may find themselves relying on] … an excessively high degree of rationality [which] threatens to trivialize the model. For the more rational subjects are, the less surprising it is that they are self-aware." (Gertler, 2003/2008 Section 2.4)

The suggestion is that it need not be surprising that a *rational enough* subject can introspect, in the way in which, according to Shoemaker, a subject with no quasi-perceptual self-knowledge can introspect. What is alleged to remain in doubt, is the proposal that normal introspecting subjects have the degree of rationality required.

This objection may seem plausible if one concentrates on the aspects of Shoemaker's presentation where he lays out the lines of reasoning which *rationalize* the transitions which self-blind subjects can make. But as has been emphasized above (3.3.3.2, 3.3.4.2 and 3.3.5.2), the transitions themselves *do not involve* following these lines of reasoning. They just involve making the relevant transition, in one step; and not even by thinking *about* that step, the step is just a rational step which certain kinds of subject can take[80].

So, we can say with little further ado that it would be a misunderstanding of Shoemaker's model to think that it requires an excessive degree of rationality, *if* this thought is based on the idea that the amount of rationality required is that involved in following the complex lines of thought, as opposed to that involved in making the single-step transitions.

In Gertler's review of such models of introspection, she would look to be at least straying close to this misreading, since it would seem to be only the more complex level of rationality which would render Shoemaker's account '*trivial*', in Gertler's sense: where it becomes essentially obvious that a creature who can run though 'all that' reasoning can work out what it is thinking.

3.6.2.2 The More Pointed Reading

Perhaps, though, the above comments misread Gertler's objection. For there is a much more pointed objection in the same area. Assuming Shoemaker's account is indeed read as it should be, there is a rather different sense in which the account *might* be alleged to be 'trivially' true, in as much as that the level of rationality required (to make the relevant introspective steps) is effectively *defined* as the level of rationality required to introspect.

If this *is* the line of Gertler's objection, however, my response would be that it is hardly fair to characterise the point as trivial. Shoemaker's arguments make us realise what is surely *not* obvious to most: that for any introspective transition[81], that transition can be made as a single rational step. This point seems far from trivial. And it is only once this non-trivial point is understood, that the further 'trivial' point comes into view: that if a creature has *this* much rationality (and no more), it can introspect.

---

[80] See esp. (Shoemaker, 1988 Section IV) on this theme.

[81] Or, at least, for all the transitions Shoemaker has discussed; but see Section 3.7 below.

Nevertheless, this rather non-trivial observation, about the relation between rationality and introspection, *does* still leave open the possibility that this type of introspection is over-rational, in a more subtle sense. Perhaps it might be the case that the level of rationality required to make such single steps, is still more than the level required simply to introspect.

Is there *any* level of rationality required, *a priori*, simply to introspect? There is. At least, there is on the view of mind as action in a space of reasons (Section 2.3). For we tell, from the third-person, that a creature has introspected when it shows signs of having its own mental states as reasons for its actions. So the minimal level of rationality required to introspect (in this public, behavioural sense) is the level required to understand[82] the deliverances of one's introspective 'mechanism'.

As such, the remaining 'over-rational' objection, is the objection that even these basic rational transitions involve more rationality than would be required simply to understand the deliverances of introspection, were it delivered in some other (for instance, quasi-perceptual) way.

How should an advocate of the rationality model respond to *this* objection? It is hard to do so within the framework of Shoemaker's arguments, for the two reasons we have just mentioned (in Section 3.6.1). Firstly, Shoemaker does not emphasize that he has a positive model of introspection. Secondly, he repeatedly states that his aim is to show that the quasi-perceptual model of self-knowledge is of dubious coherence. Therefore, if we work entirely within Shoemaker's framework of presentation, any response to the above objection must involve comparing a model which is not even presented as such, with another model which is presented as being only dubiously coherent!

In order to move this discussion forwards, we will turn next to another objection to Shoemaker's account, due to Kind (2003), which really pushes at these aspects of Shoemaker's presentation. It will be argued that, in response to Kind, Shoemaker should accept that quasi-perceptual self-knowledge is a coherent possibility. Equally, it will be argued, Shoemaker's position should be strengthened, emphasizing what is anyway true, that he has presented a positive model of introspection, not just arguments against the quasi-perceptual model.

[82] This doesn't mean 'understand' in any theory-involving sense, but it does mean 'show a basic level of rationality towards'.

Once we are clear that there are two different, coherent, types of self-knowledge in question, here, we can compare the two. On doing so, we find strong reasons for thinking that the rationality-model describes genuine, first-person self-acquaintance, and that the quasi-perceptual model does not (even though both are coherent ways of gaining self-knowledge). Equally, we find compelling reasons for thinking that the rationality model is *not* over-rational.

3.6.3 Kind's Objection

Kind (2003) offers a novel objection to Shoemaker's arguments against the quasi-perceptual model of introspection. Remember, Shoemaker argues that self-blindness is possible only if introspection is quasi-perceptual, and he also argues that self-blindness is not possible.

Kind is prepared to accept most of the steps of Shoemaker's argument. Importantly, she agrees that, if introspection is quasi-perceptual, then self-blindness is a possibility. (Remember, self-blindness is the situation of being as rational as a normal person, but lacking introspection.) Therefore, she *accepts* that, were there a successful argument against the possibility self-blindness, we would have a successful argument against the claim that introspection is quasi-perceptual.

Further, she accepts[83] that *most* of Shoemaker's line of argument against the possibility of self-blindness is successful. She accepts that his reasoning goes through, up to and including the conclusion that:

> "George is aware of his own beliefs and desires to the same extent as a normal person would be" (Kind, 2003 p.44)

That is, Kind agrees that Shoemaker has shown that someone lacking *self-acquaintance*, but who is as rational as "a normal person" (p.47), can attain exactly the same degree of *self-knowledge* as someone with self-acquaintance. I should clarify the italicised terms here. Kind's usage follows Shoemaker's, and I will follow the same usage in this discussion. *Self-acquaintance* is used synonymously with introspection in its most general sense: an ability to gain *self-knowledge* in a distinctively first-person way (but where this distinctive way is not presupposed to be quasi-perceptual) (Kind, 2003 p.40). *Self-knowledge*, on the other hand, is *not* used in its most general sense. It

[83] At least for the purposes of the specific counter-argument she presents, though some reservations are expressed (Kind, 2003 p.48).

refers to (the ability to gain) that knowledge typically gained through introspection (i.e. as Kind puts it "knowledge of one's mental states", p.41), but without any presupposition one way or the other as to *how* such knowledge was gained. Under these definitions, self-acquaintance grants all and only self-knowledge, but not all self-knowledge need be gained by means of self-acquaintance.

As such, Kind's objection to Shoemaker is that his arguments involve "a conflation between of the notions of self-acquaintance and self-knowledge" (p.42). Kind believes that Shoemaker has shown that anyone as rational as a normal person can gain *all* the same self-knowledge as the rest of us. But she does not believe that Shoemaker has shown that this "surprising" (p.47) ability amounts to self-acquaintance. Instead, all Kind believes Shoemaker has shown is that:

> "any person who is self-blind must acquire *by third-person means* the full extent of self-knowledge that those of us who are not self-blind acquire by first-person means" (p.47, emphasis added)

It is my contention that this objection misses the force of Shoemaker's arguments. However, I will claim, this misunderstanding is not that surprising, given that Shoemaker continually emphasizes that his aim is to show that the quasi-perceptual model of introspection is not truly coherent, and that he does very little to emphasize that his arguments amount, not just to arguments *against* the quasi-perceptual model, but also to arguments *for* a different, superior model of introspection.

What is my basis for the claim that Kind has misunderstood Shoemaker, rather than that she has simply presented an argument against him (which either succeeds or fails, without involving misunderstanding)? The key issue is Kind's claim that Shoemaker has shown that George (the self-blind man) has attained his self-knowledge *by third-person means* (see the quote immediately above). Kind elaborates on this point:

> "the very discussion of Moore's paradox and rationality points to exactly the sort of third-person evidence to which the self-blind person might become attuned. We are supposing that George will answer 'Yes' to the question 'Do you believe *p*?' whenever he would answer 'Yes' to the question 'Is *p* true?'. But this means that George might very well reason that whenever there is strong (or unambiguous) evidence for some claim *p*, he should form the judgement that he believes *p*. He can use the third-person evidence for *p* itself as evidence for his belief that *p*." (p.46)

This is a critical misunderstanding of the force of Shoemaker's arguments, because it misses that point that the ability of a subject to use the third-person evidence for *p* as

evidence that the subject believes *p* is an *essentially first-person ability*. To see why this must be so, note that if I want to judge whether or not *you* believe that *p* (or even whether or not you *know* that *p*), then evidence for *p* is simply not enough. I also have to have *third-person* evidence about your relation to state of affairs *p*. But in the case of myself, I do not need any third-person evidence about my relation to *p*. I might even have forgotten everything about how I came to know (or believe) that *p*. Nevertheless, *if* I turn my mind to the question of whether *p* is the case, and find myself in a state where I am willing to speak and act on the basis that *p*, then I cannot remain rational unless I also conclude that I believe that *p*.

As emphasized earlier in the chapter (3.3.5), I am not rational in self-application of the concept of belief, unless I can make this additional step correctly. Furthermore, and crucially, *no third-person evidence about myself* is involved, here. When I talk about 'finding myself in a state where I am willing to speak and act on the basis that *p* is the case', this is quite different from the case of 'finding' another agent in that state. *If* I am rational in the self-application of the concept belief, then I do not need *any* third-person evidence (e.g. observation of current or prior behaviour, or *observation* of current or prior relation to worldly states of affairs), in order to conclude that I believe that *p*. Putting things in more Shoemakerian terms, failing to be able to reach this conclusion (*without* further evidence) is a failure of rationality. Whereas, I always do need such additional evidence, in order to decide whether or not *someone else* believes that *p*.

There is nothing in Kind's paper to show that she makes the first kind of mistake which, I have suggested, it is easy to make, concerning the rationality model of introspection. Kind does not think that the model involves the subject reasoning through, step by step, using the line of argument which Shoemaker presents merely to show *why* the single-step is rational. I'm accusing her of making a different mistake: of not realising that the single-step of rationality in question is *fundamentally first-person*.

3.6.4 On the Coherence of Quasi-Perceptual Self-Knowledge

3.6.4.1 Why Take Quasi-Perceptual Self-Knowledge Seriously?

I have just suggested that Kind (like Castañeda before her; see Section 3.5.2) has misread the rationality model of introspection, and supposed that it describes a third-person type of knowledge, when it really does not.

But even if Kind were to *accept* my arguments here, she might *still* think that "those of us who are not self-blind" (p.47) acquire our first-person self-knowledge by different

means to that forced on the self-blind. She might accept that a certain kind of rationality allows us to make this step; she might even accept that this step is fundamentally first-person in nature (and, hence, a form of self-acquaintance: first-person knowledge of mental states acquired in a fundamentally first-person way); but still she might think that it is not the sort of self-acquaintance which typically occurs, in us.

As far as I can see, the best way to address this issue is to take the possibility of quasi-perceptual self-knowledge seriously for long enough to show that, even to the extent that it is coherent, it is a much *worse* candidate for self-acquaintance than is the first-person rationality of the Shoemaker-Sellars model. This involves showing that, as soon as you have the rationality required to *understand* the deliverances of a quasi-perceptual self-knowledge mechanism, then you have no need of such a mechanism.

3.6.4.2 What Would Quasi-Perceptual Self-Knowledge Be?

It is tempting to say that quasi-perceptual self-knowledge is self-knowledge mediated by knowledge of states only contingently related to the facts known. In fact, for subtle reasons, this is a caricature. To show why, and to show what it really at issue, I will firstly spell out this caricature explicitly.

If introspection involved *knowledge* of such contingent facts, then the situation would be such that if one became aware of, say, (a mental image of) a red light, one might realize that this means that one believes that Paris is the capital of France; whereas a green light might indicate belief that Reims is the capital of France. For such a view to work, then when the contingent facts which mediate one's self-knowledge vary (or seem to), one's 'introspective' judgement must vary too. Otherwise, access to these contingent facts is *not* mediating self-knowledge in the way supposed.

Even on the 'mental image' reading, this view is a caricature of the most appealing quasi-perceptual view. At issue is whether or not the subject must have *knowledge* of the 'internal' (contingent) facts. We arrive at a position where we are taking such views seriously, if we change the caricature only in this respect: what the subject needs is *acquaintance* (in the sense discussed in Chapter 2, footnote 41 and in Section 5.6) with the contingent facts, not *knowledge* of them.

Thus, when I introspect that I believe that Paris is the capital of France (or, as it might be, that I am now seeing a blue square) it should not be supposed (even on the quasi-perceptual model) that I do this by *knowing* some contingent (only contingently related to what I eventually introspect) mental facts. Nevertheless, the quasi-perceptual model

does suppose that I come to know mental facts in virtue of *some* kind of acquaintance with facts which are only contingently related to what I eventually know. This is what makes the model quasi-perceptual: I come to know mental facts about myself *by* coming to know (or, at least, *by* being somehow acquainted with) some more specific facts which are only contingently related to the public mental facts.

The trouble with such views, according to all of Shoemaker's arguments, is that once I am in a position to understand the deliverances of such a quasi-perceptual mechanism, *then I no longer need it*. The rationality required simply to understand the deliverances of quasi-perceptual introspection is already enough to be able to introspect in a different, but perfectly first-person, way.

To see why this is so, take the Moore-paradoxical example. Remember that Shoemaker's arguments run as follows: if one is prepared to endorse (even in thought) Moore-type sentences (or thoughts) then one has not understood *what it means* to believe something, at least as far as the concept of belief applies to oneself. On the other hand, if one is rational in self-application of the concept of belief, then one no longer *needs* a quasi-perceptual mechanism, mere rationality is enough to arrive at the same knowledge (this is something which Kind has already conceded).

Now, let's look at how this line of argument works against the caricatured quasi-perceptual model. Shoemaker's arguments show that, once one is in a position to truly understand what the 'coloured lights' (or mental images of them) indicate, then one is already in a position to know whether or not they are indicating correctly. That is, one is already in a position not to *need* access to these indicators, merely by being in a position to make use of them.

Indeed, the position is worse than this, for the self-knowledge attained by 'mere' rationality is *more authoritative* than any self-knowledge gained on the basis of these 'internal lights'. For it follows from the fact that one doesn't need these indicators, that one is also in a position to *know when they go wrong*[84], merely by being in a position to understand them.

The problem for any quasi-perceptual account of introspection is that the same points follow through for the non-caricatured position: for the quasi-perceptual account to make sense at all, it has to be supposed that when the contingent facts with which one is

[84] 'Going wrong' either by green lighting up when red should, *or* by the detector detecting green when it should detect red.

acquainted vary (or 'seem' to, if one is mistaken) then the quasi-perceptual judgements will vary. But in the same way as on the caricatured view, as soon as one is in a position to truly *understand* one's quasi-perceptually based judgements, one is already in a position to know whether they are right or not, *merely* by exercise of one's rationality. One know longer needs any quasi-perceptual mechanism.

As far as I can make out, the above lines of argument are entirely Shoemakerian in spirit. The difference from Shoemaker, in detail, is that I have allowed that there can be quasi-perceptual self-knowledge, in order to show why it is not needed. That is, I believe we can be clearer than Shoemaker is, about the problem with quasi-perceptual self-knowledge. It is *not* incoherent merely to suppose that we have such a means of access to our mental states. We could literally have such access. There could, in principle, be lights (even external, physical lights) wired up such that they go green for one belief, and red for another. The lights could even be fairly reliable, and I could understand what they indicate when they are working. What is not coherent, if the above arguments are right, is the claim that I might be in the position of *having* to use such a mechanism (on either the *knowledge* or the *acquaintance* model of quasi-perceptual introspection) in order to introspect.

3.6.4.3 What Mechanism Needs To Evolve, For Introspection?

It is interesting to note that the above arguments follow through even to the mechanistic level. A creature with a physical organisation sufficient to enable it to act in the manner required for it to be said (from the outside) to show evidence of introspection is already a creature which has no need for some *further*, quasi-perceptual mechanism. For it cannot be coherent *both* to suppose that the creature can *understand* its introspective thoughts (i.e. have them at all, on the model of mind presented in 2.3) *and* to suppose that it gains this knowledge in a way which can go wrong as it could if the creature were relying on some further mechanism, *above and beyond what is required merely for it to be said to have such understanding*.

Returning to a point made above (Section 3.3.2), about the plausibility of the rationality model of introspection, I think it is often supposed (by objectors to the model) that the quasi-perceptual account requires *less* of a physical system, not more. This is false. There would always be *more*, not less, which would need to be evolved, in order for a creature to have quasi-perceptual 'introspection' (as opposed to direct, rational introspection). There would have to be the understanding, and then some

additional 'detection' mechanism which could go wrong, even when this understanding was working perfectly. This 'more', this extra mechanism, would therefore always be entirely redundant: for a physical system which instantiates the minimal possible rational understanding required to make use of a quasi-perceptual mechanism already has no need of such a mechanism.

I believe it cannot be successfully countered that the quasi-perceptual mechanism is a *part* of the rationality in question. For something is not a quasi-perceptual mechanism at all, unless there is a possibility of its failing (misdetecting the relevant inner states) whilst the creature can still fully (i.e. as well as before) understand the erroneous deliverances. Shoemaker's arguments show that anything which can fail in this way *will always be surplus to requirements.*

3.6.4.4 Is Quasi-Perceptual Self-Knowledge Really Introspection?

There is a final, related, point to make. Kind quite correctly characterises introspection (self-acquaintance) as follows:

> "What makes self-acquaintance special consists in the fact that no one but I can acquire knowledge via this sort of access to my mental states" (Kind, 2003 p.40)

But, I will now argue, by this very criterion, quasi-perceptual self-knowledge has less claim to be counted as self-acquaintance than does the single-step, rational self-knowledge which Shoemaker describes.

Once again, I will firstly make the point using the caricature of quasi-perceptual self-knowledge. We can see straightforwardly that it is quite possible that I could be aware of 'lights' or 'indicators' which make me aware of *someone else's* mental states. As such, *there is nothing about this way of coming to acquire knowledge of mental states which makes it intrinsically first-person.*

As before, the very same point applies to the non-caricatured model of quasi-perceptual introspection. For, *whatever* the acquaintance is, which I have with states which are only contingent indicators of my public mental states, it is perfectly possibly that I could have *acquaintance of the same type, with states which are still my internal states*, but where the states in question have now (due to some deviant causal chain) been modified such that they indicate *someone else's* mental states.

There some could be some non-standard 'wiring', and a radio-enabled link between subjects, such that the quasi-perceptual mechanism in me now could be changed, at the flick of a scientist's switch, to give me information about (or, acquaintance with states

which give me information about) *someone else's mental states*. Of course this would be confusing at first, but a subject could perfectly well come to master this situation, and to be aware of someone else's mental states *in this way*. Indeed, if the quasi-perceptual model is coherent then even the subject in the initial state of confusion would *in fact* be gaining awareness of someone else's mental states, they just would not yet understand that this was so.

As such, quasi-perceptual self-knowledge is *not* a fundamentally first-person mode of gaining knowledge *of mental states*, even though, in its most plausible form, it *does* involve fundamentally first person knowledge of (acquaintance with) the 'internal' states which indicate those mental states. Someone else can, in principle, acquire knowledge of *my* mental states by acquaintance with *their* internal indicators, and *vice versa*.

However, someone else can *never* acquire knowledge of my mental states in virtue of making a one-step rational transition in my mind. This must surely follow on almost any reasonable model of mind: how can someone else make a rational transition *in my mind*? It certainly follows if we adopt the space of reasons model of mind outlined in Section 2.3. On that model, if there were two physically separate agents (or, rather, physical bodies) which were so closely linked that they shared one and the same space of rationality, then they would share a mind, on the space of reasons definition. In a sense, there would only be one agent, with an uncommon bodily form. That (uncommon) agent would be introspecting, but it wouldn't be a case of one agent knowing another agent's mind.

All of this emphasizes what Shoemaker has not emphasized: how good a claim he has to have presented a different, *better* model of introspection than the quasi-perceptual model. For introspection on Shoemaker's model is a deeply, fundamentally reflexive act by a subject – exactly as it should be. It is *strictly* true, on the rationality model of introspection, that (as Kind requires) "no one but I can acquire knowledge via this sort of access to my mental states" (Kind, 2003 p.40). However, on the quasi-perceptual model, *even if* it is strictly true that no one but I can have first-person acquaintance with the 'internal' 'features' which contingently indicate mental states, it is *not* true that the mental states which those features indicate have to be mine.

For this reason, I would argue, not only is quasi-perceptual introspection never *required*, in order to introspect, it is not even truly introspection[85].

3.7 Why Shoemaker's Account Generalises

In this penultimate section, I wish to briefly argue that Shoemaker's account generalizes. That for any property defined by its role in a space of reasons, that property can be introspected by a sufficiently rational creature, without the need for any additional, quasi-perceptual mechanism for detecting internal states.

Recall Shoemaker's arguments concerning the introspection of belief and desire. These arguments work because belief and desire are *defined* by their role in a space of reasons.

Recall the case of third-person ascription of such a state. It can never be rational to believe that some other subject is behaving in the way constitutive of believing *x* (say) (and under that description) and yet not to believe that the subject believes *x*. One has not fully understood the concept, if one refuses to make this transition.

Now recall the case of first-person ascription of such states. Because these states are defined by their role in rationality, it must always be true that the ability to self-ascribe such a state in a single step is no more nor less than an exercise of rationality. Conversely, the failure to self-ascribe such a state (in a single-step) must amount to the failure of rationality as regards self-ascription of the mental concept in question. (Nothing guarantees that a creature must have such rationality, of course.)

That is always the structure of Shoemaker's arguments. This is why he struggles to extend the arguments to pain (Section 3.3.6), for he remains of the opinion that pain cannot be fully characterised by its role in rationality (Chapter 4).

But that as it may, for now we can note that the above argument form will always be available, for any property which *can* be fully defined by its role in rationality.

This is not to say that any such property will be introspectible in a creature which has it. What Shoemaker's arguments do (and perhaps do better in combination with the observations above in 3.6.4 about quasi-perceptual self-knowledge) is to show *what it would be* to be able to introspect such a property. There can be no prior legislation as to

---

[85] Despite the deeply misleading use of the word 'introspection', in that case. However there is widespread agreement, in work on this issue, that although the term 'introspection' etymologically begs the central question, it can be used (for the purposes of such discussion) in a non-question begging, general sense (with the same meaning as self-acquaintance as defined above, in Section 3.6.3).

how much rationality a creature has; in particular, Shoemaker explicitly allows (as would I) that a creature can be rational enough that some mental concept applies to it, without possessing the additional level of rationality required to introspect and know that the concept applies to it (Shoemaker, 1988 Section III).

3.8 Introspection of Intrinsic Properties

If this account of introspection is correct (and it certainly seems physically possible, and compelling, once one realizes that the possibility exists) then one could not find out anything intrinsic about one's being in introspection, be it physical ('60 Hz neural firing') or mental (the strong phenomenal realists' qualia). More precisely, what one comes to know, in introspection, would legitimate nothing more specific than a conclusion such as: 'I am a physical agent structured in some way (I know not which) which is compatible with my behaviour (and counterfactual behaviour) having the structure I have just introspected' (for instance: compatible with the fact that 'I am now perceiving *x*', or compatible with the fact that 'I believe that Paris is the capital of France').

However, it is still widely supposed that such an account of what we come to know, by introspection, is false. Specifically, it is very widely supposed that, in coming to know the phenomenal or qualitative aspects of our mental states, we are coming to know something which *is* more specific, in a way contrary to what I have just outlined.

Indeed, this is exactly the point of the *a posteriori* accounts of phenomenal knowledge canvassed in Chapter 2. For if (say) inverted spectra are possible, then I *do* have more specific knowledge of this problematic type. I know that I have *this* phenomenal feel, rather than *that* one. The phenomenal feel is not fully specified, by specifying the public mental state.

Surprisingly, in the next chapter, we will see that Shoemaker's own current account of qualia has not fully escaped the notion that we can come to have such *a posteriori* knowledge when we introspect our own qualitative states, and we will present reason to think that Shoemaker must be wrong about this – even by his own lights.

4. Shoemaker's New Account of Qualia

4.1 Introduction

If the model of introspection described in the previous chapter is correct (as I believe it is), then introspection gives us access only to properties at what Lewis describes as the *a priori* mental level (see Section 2.2.4). This is entirely in accord with the line of argument developed in Chapter 2, in which it was argued that any other kind of introspective knowledge would threaten the possibility of an eventual naturalisation of the mental.

On the rationality model, in introspection we do not come to know any 'more specific' fact about our mental states than the public mental facts such as 'I am now seeing a red rose'. This stricture on what we can know in introspection has two aspects. Firstly, we cannot come to know anything more specific (c.f. Sections 3.3.8 and 3.8) about what *physical* state we are in. Secondly, we cannot come to know anything more specific about which *mental* state we are in, in this very particular sense: we cannot come to know any (allegedly) mental fact which is not pinned down, by pinning down the *a priori* mental level.

I argued in Chapter 2 that the preservation of physicalism required this result. In Chapter 3, I have just presented an independently plausible model of introspection which itself entails this same result.

However, there is a problem. The main proponent of this very model of introspection, Shoemaker, is himself unwilling to accept this result. In his more recent work on qualia, Shoemaker continues to endorse the claim that our intuitions about qualia can only be resolved by allowing a certain *a posteriori* aspect to what we know, when we introspect:

> "I think that the best response to these worries [Jackson's Knowledge Argument and the apparent possibility of inverted spectra] is to show that the existence of these apparent disparities between manifest and scientific image is just what an acceptable physicalist theory should lead us to expect. What I shall argue is that a broadly functionalist view, combined with physicalism, predicts that we will be presented in experiences with a phenomenal character that is in a certain sense irreducible to its functional and physical underpinnings." (Shoemaker, 1994c Section IV, p.261).

Indeed, as the above quote suggests, the main aim of Shoemaker's most recent work on qualia is to continue to argue that the intuition behind the inverted spectrum is correct, in precisely the phenomenal concept strategist's sense: that the public mental facts are not enough to pin down the qualitative mental facts. However, as Shoemaker himself says:

> "there is a prima facie strain between my opposition to the object-perception model of introspection and my being what Frank Jackson calls a 'qualia freak.' " (Shoemaker, 1994d p.21)

Nevertheless, Shoemaker believes that he can resolve that (at least *prima facie*) tension. Because of this, I might well seem to be caught on the horns of a dilemma. Either I have misdescribed Shoemaker's model of introspection or, if I have described it correctly, I have misdescribed its consequences for self-knowledge. Neither of these, of course, would be good news for the approach being developed in this thesis. In fact, there is not a dilemma here, but a trilemma. The third, and *prima facie* less plausible, option is that Shoemaker himself has not fully embraced the implications of his own model of introspection for self-knowledge of qualitative mental states. It will be the burden of this chapter to argue for this third outcome.

It turns out, though, that this chapter is important for another reason, too. I find that although I disagree with the most fundamental details of Shoemaker's new model of qualia, I am nevertheless strongly persuaded that many of the only slightly less fundamental details are correct. As such, in reading the presentation below, it should be borne in mind that the aspect of Shoemaker's account of qualia which I reject is its inclusion of properties (Shoemaker's new qualia) whose existence and/or nature remains *a posteriori* with respect to the public mental level. That is what I reject. Shoemaker also says much about the relation of qualia to the normal, public properties which we perceive; and I am now convinced that much of what he says about *this* should be preserved[86]. I therefore find myself persuaded to include *these* aspects of Shoemaker's model of qualia in the account which I will present in the next chapter.

---

[86] This is perhaps doubly strange, since what Shoemaker says in this regard *appears* to have been forced on him as a kind of rearguard defence of *a posteriori* qualia. A detailed analysis of exactly why he (and now I, influenced by him) find these shared features compelling, despite such apparently fundamental differences in approach, is unfortunately beyond the scope of the present thesis.

Bearing in mind that I find myself in the strange position of arguing against the single most fundamental aspect of Shoemaker's account of qualia, but arguing for many only slightly less fundamental aspects, I can proceed to presenting, and then (at least as regards the *a posteriori* aspect) arguing against, Shoemaker's new model of qualia.

4.2 Why Shoemaker Needs a New Account

4.2.1 Introspection of Perceptual Contents

I have argued that it follows, from Shoemaker's account of introspection, that we can only introspect our own public mental relations to (at least seeming) public properties. In his recent work on qualia, Shoemaker himself seems explicitly sympathetic to this. He says:

> "if asked to focus on "what it is like" to have this or that sort of experience, there seems to be nothing for one's attention to focus on except the content of the experience" (Shoemaker, 1994c p.257)

The statement might sound tentative, but in fact Shoemaker does mean to endorse this claim, or something very like it, as we will see below. However, if there really is nothing else for us to focus on except the content of experience, this would seem to leave no room for introspectible qualia. If this claim is correct then, when introspecting the content of our perception, we can *only* discover what there is (or seems to be). It would seem that we cannot introspect any aspect of *how* what there is (or seems to be) is presented.

We should note, firstly, that the above claim about what we can "focus on" is not quite correct, even on Shoemaker's account of introspection. In introspection, we can indeed know what we perceive, or seem to[87], but we can *also* come to know the mental relation which we have, to what there seems to be (for instance, that we are *seeing* state of affairs *x*). Presumably Shoemaker sees these mental relations, which we can also come to know in introspection, as irrelevant to the issue of naturalising qualia. This will be questioned in the next chapter. Why should Shoemaker find these properties irrelevant? Because, for him, qualia (or, at least, phenomenal properties – see below) must be properties which *appear* to us in the contents of perception. As he says:

---

[87] What Evans' means, when he talks of a subject who "gazes again *at the world*" (Evans, 1982, original emphasis). In many ways, Evans' brief discussion of introspection is very reminiscent of the Shoemaker-Sellars position (see also footnote 90).

> "what seemed to pose the problem was the experienced character of redness, sweetness, the sound of a flute, and the smell of a skunk. And *these* are not experienced as features of sensations or sense-experiences; they are experienced as features of things in our environment." (Shoemaker, 1994d p.25)

The problem, for Shoemaker, is that there would seem to be nothing in the public contents of perception which matches the supposed subjectivity of such features. In this chapter, we will look at Shoemaker's attempt to find (relational) properties in the contents of perception which can naturalise the subjective aspect of such public features. First, though, an important clarification as regards what exactly we are supposed to be looking at, when we are looking at 'the contents of perception'.

4.2.2 Content or Contents?

I have just quoted Shoemaker's claim that, in introspection, we can only "focus on … the content of the experience". But what, exactly, does Shoemaker mean by content? This is pertinent because there are two different usages in the literature, and it is often only via context that one can tell which is in play.

One is what we might call the 'Oxford' usage. In labelling it thus, I am thinking of authors such as Peacocke, McDowell, and Evans – though, more broadly, this usage is common to most of those authors involved in the nonconceptual content debate, on both sides of the Atlantic (for a little more on this debate, see the Appendix). To a good first approximation, when used thus, content is identified with Fregean sense. It is the mode of presentation of some referent. In that case, to be aware of the content of an experience would be, to be aware (presumably, under some further mode of presentation) of a mode of presentation.

The other usage is what one might, more broadly, call the 'American' usage. On this usage, the 'contents of perception' refers, in the first instance, to *what* is perceived, rather than *how*. Certainly, on this American usage, Fregean sense (or something playing a similar theoretical role) is still relevant: for content understood thus is always present to a subject under some mode of presentation. Nevertheless, in this usage, content doesn't *mean* mode of presentation; instead, it means that which is presented, or seems to be, in having the experience.

This is the meaning which Siegel endorses as being most fundamental, with her introductory remarks to the *Stanford Encyclopedia of Philosophy* article on '*The Contents of Perception*':

> "In contemporary philosophy, the phrase 'the contents of perception' means, roughly, what is conveyed to the subject by her perceptual experience. For example, suppose you are looking into a piano at the array of hammers and strings. There will be a way these things look to you when you see them: they will look to have a certain shape, color, texture, and arrangement relative to one another, among other things. Your visual experience conveys to you that the piano has these features. If your experience is illusory in some respect then the piano won't really have all those features; but even then, there will still be something conveyed to you by your experience." (Siegel, 2005/2008)

This is typical of what I am calling the 'American' usage: the clear emphasis, in the above, is on *what* is presented (despite the at least implicit importance of *how*). 'Contents' in the above are the (at least intentional) objects and properties which one sees (or seems to), not (as in the 'Oxford' usage) the *way* in which such (at least seeming) properties are presented[88].

One might at first think that content, on what I have called the American usage, means the same as Fregean referent. But of course we can have an illusion as of a ripe tomato, just as we can have a veridical experience of a ripe tomato. There is only a Fregean referent in the latter case, though there might seem to the subject to be, in the former. However, the intentional content in both cases is 'the ripe tomato', the (perhaps merely intentional) object of the experience.

It is this latter usage which Shoemaker is adopting. He likewise adopts the venerable example of the ripe tomato, in saying:

> "it doesn't matter, to the "what it is like" question, whether the tomato one sees is really there or is merely an intentional object. If one is asked to focus on the experience without focussing on its intentional object, or its representational content, one simply has no idea of what to do." (Shoemaker, 1994d pp.30-31)

This latter quote also helps to support my claim that Shoemaker is more than merely tentative, in his endorsement of the substantive claim that we *must* focus on the content of experience (in the 'American' sense), at least in the first instance.

---

[88] It turns out that one can get *some* clue as to which usage is in play, simply by looking at whether an author tends to use 'content' (*typically* the 'Oxford' usage), or 'contents' (*typically* the 'American' usage), however this singular/plural difference is by no means a definitive guide (as is indeed exemplified by the quote from Shoemaker which introduces 4.2.1).

4.2.3 Shoemaker's Dilemma

Shoemaker has argued against an object-perceptual model of introspection. It is because of this, that he (quite correctly) makes statements such as those quoted above, as regards *what* we can focus on, in introspection. However, as quoted in 4.1, and as Shoemaker now explicitly acknowledges, there is an at least *prima facie* tension between his endorsement of this model of introspection, and his being a "qualia freak". What, in more detail, is that nature of this tension? Shoemaker notes that no particular tension would arise, if we were only able to introspect "such intentional states as beliefs and desires" (Shoemaker, 1994d p.21):

> "[these states] include few if any of the "intrinsic" properties which, on the object-perceptual model, objects of perception ought to be perceived as having." (Shoemaker, 1994d p.21)

But, Shoemaker suggests, things seem very different in the case of "sensations, feelings, and perceptual experiences" (Shoemaker, 1994d p.21):

> "While a few philosophers have recently maintained that the only introspectively accessible properties of these [states] are intentional ones, I think that the majority view is that that these have a "phenomenal" or "qualitative" character that is not captured simply by saying what their representational content, if any, is." (Shoemaker, 1994d p.22)

He continues:

> "It is commonly held, and has been held by me, that the introspectible features of these mental states or events include non-intentional properties, sometimes called "qualia"" (Shoemaker, 1994d p.22)

We can start to see the problem. Shoemaker is happy to acknowledge that his arguments against the object-perceptual model rule out the introspection of "intrinsic" properties. All the same, he himself previously claimed that qualia are exactly such properties. The former work Shoemaker refers to ("has been held by me" in the above quote) includes papers such as his (1975) and (1982). His aim, in his more recent work on the nature of qualia (1994c; 1994d) is to modify his former view, and to hold that when we introspect, we do not discover any intrinsic properties of experience (as we cannot, if his arguments against the object-perceptual model of introspection are correct).

Nevertheless (and perhaps surprisingly), Shoemaker still wishes to hold that there *are* intrinsic properties of experience, qualia, which determine what it is like to have the experience. How can this be? How, for instance, can we ever come to know that we

have these qualia, if we cannot ever know them in introspection? As will be explained in more detail below, Shoemaker squares this circle by arguing that whilst intrinsic qualia are not directly introspectible, we can nevertheless know of their existence, given a certain theoretically informed understanding, based on what we can introspect[89]. I believe that this move, too, can and must be rejected. I argued in the previous chapter (Sections 3.3.8 and 3.8) that it follows from Shoemaker's account of introspection that intrinsic properties can neither be known directly in introspection, *nor inferred from* the properties known in introspection. Since Shoemaker still thinks that the latter, at least, is possible, I need to be very clear about why I think he is wrong. I present the central problem in Section 4.6.1, below.

For now, though, we might ask why Shoemaker feels the need to make this move at all. Would it not be simpler for him to revise his view, and to accept that intrinsic mental properties (i.e. those properties which cannot be captured in a publicly verifiable description of our at least counterfactual behaviour) are no part of our mental life? Shoemaker feels he cannot take that step, for the following reason.

4.2.4 Do We Still Need Intrinsic Qualia?

Shoemaker finds it evident that:

> "reflection on the disparity between the manifest and the scientific images makes inescapable the conclusion that, to put it vaguely at first, the phenomenal character we are confronted with in color experience is due not simply to what there is in our environment but also, in part, to *our* nature, namely the nature of our sensory apparatus and constitution." (Shoemaker, 1994d p.24)

For my part, I also find this intuition very compelling. Trivially, as Shoemaker points out:

> "At the very least, the way things appear to us is determined in part by limitations on the powers of resolution of our sensory organs." (Shoemaker, 1994d p.24)

---

[89] N.B. This is jumping ahead, however I should clarify that whilst Shoemaker's new model indeed requires a certain theoretical understanding in order for a subject to know *qualia* (as explained in Section 4.6.4), it does *not* require any kind of understanding, of the *phenomenal properties* which he introduces (of which more below), in order for a subject to see them. It is this latter issue which will prove to be a key problem with Shoemaker's account (see Section 4.6; and c.f. the replacement account offered in Sections 5.3.4-5.3.6).

That much is surely not in doubt. But, like Shoemaker (and unlike Dennett, 1988; or Dennett, 2005b, for instance), I feel that there is something more to be accounted for here, in one way or another: there does seem to be a redness to my reds, which need not be the redness of your reds, even if we both agree exactly on which things are red. Shoemaker puts it thus:

> "The intuition that this is so finds expression in the inverted spectrum hypothesis – it seems intelligible to suppose that there are creatures who make all the color discriminations we make, and are capable of using color language just as we do, but who, in any given objective situation, are confronted with a very different phenomenal character than we would be in that same situation, and it is not credible that such creatures would be misperceiving the world." (Shoemaker, 1994d p.24)

I am no fan of the classical (behaviourally undetectable) inverted spectrum. In Chapter 2, I argued that it cannot be made compatible with physicalism (at least, not in a way which does justice to the existence of a self-standing mental level). Nevertheless, it is the full-blown, behaviourally undetectable inverted spectrum which Shoemaker still means to invoke. More accurately, he means to invoke the claim that there is nothing in the concepts involved in describing the inverted spectrum scenario which rules it out *a priori*; even though he would accept, I think, that we might learn something empirically (i.e. *a posteriori*) which makes us doubt its possibility. The quote from Shoemaker given in the introductory section of this chapter helps to confirm that this is still his view, as do all the details of his revised account of qualia.

All of this, however, makes the specific wording Shoemaker which has chosen, to describe what "seems intelligible" about 'the inverted spectrum hypothesis', rather interesting, precisely because he *doesn't* in fact pin down a behaviourally undetectable case: two creatures who use language exactly as we do, and discriminate just what we do, *need not* behave the same as each other (or as us). Almost trivially, one agent might love the colour blue, and one might loathe it; one might be reminded, by red, of blood and pain, the other of celebration and good luck. These differences would certainly amount to (at least counterfactual) behavioural differences (as discussed in Section 2.2.7 and Chapter 5).

To recap, I have argued that the intuitions of strong phenomenal realism cannot be made compatible with a normal scientific explanation of qualia (Chapter 2). Moreover, one upshot of Chapter 3 would seem to be that such intuitions cannot be made compatible with Shoemaker's own model of introspection, which I have endorsed. Yet

Shoemaker still holds these problematic intuitions. He recognises a *prima facie* tension, here. But he believes he can square this circle. Now we will look in detail at how he proposes to do this.

4.3 Shoemaker's New Account

4.3.1 Projectivism

Our problem is the apparent possibility of qualitative difference, between subjects who are sensing the same objective property. One approach to this problem is to identify that which differs as something which is, and seems to be, inner: *is* inner because it is physically internal to the subject; *seems* to be inner because, in one sense or another, one turns something very like a 'glance' inwards[90], in order to discern it. As Shoemaker now realises, this is not compatible with his own account of introspection.

There already exists, in the philosophy of perception, a different approach – which is to suggest that although that which differs *is* inner, it at least *seems* to be outer. This approach is projectivism, and it has many of the features of Shoemaker's present solution to the problem of qualia. Indeed, Shoemaker introduces his solution by way of first introducing this theory, and I will do the same here.

Projectivism proposes that, as between two spectral inverts, what actually differs between them is internal to each of them, but what *seems* to differ, from their points of view, are the properties of things in the world. To one, there seems to be phenomenal blue out there; to the other, in the same place in the world, there seems to be what the first would identify as phenomenal green. This is so, even though both are picking out, say, the same surface reflectance property, and can agree on a shared word for it.

Shoemaker identifies two varieties of traditional projectivism, literal and figurative (Shoemaker, 1994c pp.250-251; Shoemaker, 1994d pp.25-26), neither of which he is prepared to accept, in their existing forms.

Literal projectivism proposes that, although qualia *are* intrinsic features of our experience, they *seem* to be features of the objects of the world: the green which I seem to see, on the leaves of the tree in front of me, is in fact some intrinsic property of my current mental state – projected, as it were, onto the leaves. Shoemaker (rightly, I think) doubts whether it is possible for a property of an experience, as such, to so much as

[90] I take this wording from Evans, another author who has argued that "we continue to have no need for the idea of the inward glance" (1982 p.226).

seem to be a property of the surfaces of things. There appears to be a type-mismatch here. As Shoemaker puts it, "an experience is an experiencing" (Shoemaker, 1994d p.25), something which happens to a subject. How can some property of *that* seem to be an extended property of external objects in physical space?

Figurative projectivism, on the other hand:

> "concedes that qualia, understood as properties of experiences, are not properties that could even seem to us to be instantiated in the world in the way in which colours, for example, are perceived as being" (Shoemaker, 1994d p.25).

Instead figurative projectivism proposes that:

> "associated with each quale is a property that can seem to us to be instantiated in the world in this way – and that when an experience instantiates a quale, the subject perceives something in the world as instantiating, not that quale itself, but the associated property." (Shoemaker, 1994d p.25)

Note that the view is still *projectivism*, inasmuch as that the property which seems to be in the world before us isn't in fact in the world before us, but is (metaphorically) projected from our experience. But in this case, the apparent properties before us turn out to be properties which *nothing* ever actually has (neither experiences, nor worldly objects) – they are merely intentional; properties which some things seem to have (Shoemaker, 1994d p.26).

Again, I'm sympathetic to Shoemaker's response to this view. He accepts that we *can* imagine properties being represented in our experience, which are never instantiated in anything; the property of being a ghost, for instance. But to do so, we have to have *some* idea of what it would be for this property to be instantiated. Thus, as Shoemaker puts it "we at least have some idea of what would *count* as someone veridically perceiving an instantiation of the property of being a ghost" (Shoemaker, 1994d p.26). However, in the case of those properties which figurative projectivism says seem to be instantiated, nothing does instantiate them, and we can have no idea of what could even count as something's *actually* instantiating them (rather then just seeming to). This throws doubt on the idea that we can truly make sense of such properties even seeming to be instantiated and, hence, of figurative projectivism itself.

Moreover figurative projectivism, like literal projectivism, is an 'error theory' – claiming that we are permanently and fundamentally misled, in our perception of the world; that we always see things that aren't, and couldn't be, there. Such theories are

often (I think rightly) taken to be unattractive for that very reason (Levine, 2003 p.73; c.f. Shoemaker, 1994d p.25, p.27).

Despite all this, Shoemaker sees many attractive properties in projectivism. Perhaps most specifically, that it allows that "we focus on the phenomenal character by focusing on what the experience is *of*" (Shoemaker, 1994d p.26) (that is, on the intentional content of the experience, in Shoemaker's terminology). Therefore, as Shoemaker puts the point, in passing, in a different paper, "I am in the uncomfortable position of funding the view [projectivism] both plausible and unintelligible" (Shoemaker, 1991 p.139, n.4)!

4.3.2 Shoemaker's Proposal

Shoemaker's new view is not projectivism. In his new view, we do not 'project' out, from our experience, onto the world, properties which aren't and couldn't be there. But neither do we look inwards and find qualia. Instead Shoemaker argues that he has identified a *bona fide*, relational, property of coloured things, which varies as required, as between spectral inverts, and which bears much the same relation to the intrinsic qualities of our experience as the 'impossible', only-apparently-external properties which projectivism proposes.

Shoemaker sees it as a strength of projectivism, that that which varies, as between spectral inverts, is in the intentional *contents* of experience (i.e., in the 'American' usage of 'contents'; in that which there seems to be). That is, irrespective of arguments from introspection, he now sees this as a strength, over and above views such as his earlier view, which attempt to locate that which seems to differ in some 'internal' or intrinsic property of the subject. Why should it be more appealing to locate that which varies in the properties which we see before us? As Shoemaker puts it (repeating a quote already given in 4.2.1):

> "what seemed to pose the problem was the experienced character of redness, sweetness, the sound of a flute, and the smell of a skunk. And *these* are not experienced as features of sensations or sense-experiences; they are experienced as features of things in our environment." (Shoemaker, 1994d p.25)

So now, in order to resolve the tensions in his account (at least, as he believes) Shoemaker proposes that qualia are *non*-introspectible, but still intrinsic, properties of experience. He further proposes that if the quale associated with the public property red in some subject is *R*, then what that subject perceives, what varies between spectral

inverts, is what he calls the *phenomenal property*, *R**: the property of producing (or being disposed to produce) experiences with quale *R*.

To see how this works, Shoemaker asks us to imagine a situation in which *Q1* is the quale associated with redness in Jack, and *Q2* is the quale associated with redness in Jill. Then red has the property of producing, or being disposed to produce, Q1 in Jack; and it also has the property of producing, or being disposed to produce, Q2 in Jill. As Shoemaker says, the public property red indeed has both these relational properties, and neither of these properties is the property of being red (Shoemaker, 1994d p.27).

Here is Shoemaker's proposal as to how all this relates to qualitative character:

> "In all of these cases [of experience with phenomenal character] the phenomenal character of the experiences consists in a certain aspect of its representational character, i.e., in its representing a certain sort of property of objects, namely "phenomenal properties" that are constitutively defined by relations to our experience." (Shoemaker, 1994d p.31)

That is to say, the phenomenal character of my experience of red consists in the fact that my experience represents (has as intentional object) the relational property *R** (the property of producing or being disposed to produce a certain quale, *R*, in me).

Some clarifications are certainly in order. Firstly, one might worry that neither Jack nor Jill actually see the tomato as red, if each sees it as *Q1** and *Q2**, respectively. But note that *Q1* and *Q2* are, indeed, relational properties of red. Shoemaker's proposal is that "experience represents color *by* representing the phenomenal property" (Shoemaker, 1994d p.35), and also that "these two properties are conflated in the content of the experience" (Shoemaker, 1994d p.35). In his opinion "the view that there is this two-fold character to the representational content of experiences is prima facie counterintuitive. But the alternatives are much worse." (Shoemaker, 1994d p.35). Reading on, the 'worse' alternatives include the variants of projectivism mentioned earlier (I agree that this *is* worse), but they also include those views which have "no explanation to give of the seeming discrepancy between the world as we experience it and the world as science says it is" (Shoemaker, 1994d p.35). Of course, it is precisely Shoemaker's certainty that there *is* such a seeming discrepancy which requires him to allow intrinsic qualia (properties not capturable at the public mental level).

To my mind, Shoemaker's new account does get something right. For, *to the extent that there is qualitative, subjective character to experience at all* (even if, as I will claim, that character can be analysed *non*-intrinsically), then the things we see clearly do *have* relational properties of something like the sort Shoemaker describes – whether

or not we *perceive* them to have such properties. That is, if there is any sense to be made of the claim that my experience of red has a certain qualitative character which might be different from that of your experience of red, then it certainly follows that red has the property of producing, or tending to produce, this qualitative character in me and that qualitative character in you. Equally, *given the concept of qualitative character*, it is perhaps also not implausible to propose that we can see, or learn to see, things *as having* these relational properties (see Section 5.3.6).

But this is not all that Shoemaker is claiming. He is claiming that the least theoretically informed of us sees every colour which we see, in and by having experiences which represent (have as intentional contents) such relational properties. Can this be right? There are various objections to such a view. Some of these, Shoemaker handles perfectly well. I will mention a few of these responses in Section 4.4 as they help to clarify Shoemaker's view – and, perhaps, to show why *some* aspects of it are appealing. But there are other problems which cannot be so successfully handled, or so I will argue in Sections 4.5 and 4.6 (where we will see that Shoemaker has been honest enough to acknowledge, if not to elaborate upon, some of the key areas of tension in his view).

4.4 Some Resolvable Issues

4.4.1 Can 'Phenomenal Red' be Relational?

One line of objection builds on the observation that we have no intuitive sense, in our experience of colour, that we are experiencing a relational property. Hence, this objection goes, Shoemaker's analysis is phenomenologically inapt. This objection, at least, I think Shoemaker addresses perfectly well. He asks us to consider the property "*to the right of*" (Shoemaker, 1994d p.28). He suggests (and this seems correct to me, though it is purely an intuition) that pre-theoretically, we experience 'to the right of' as a dyadic relation: *A* is to the right of *B*. But, as Shoemaker rightly points out, reflection reveals that "*to the right of* is, at least, triadic" (Shoemaker, 1994d p.28). *A* may be to the right of *B* from my point of view, but might be to the left from yours. (Shoemaker is also right to say "at least" triadic; we don't just need a point of view, we need a point of a view and a defined up-down axis, before we can start to operationalize a relation with something like the properties of 'to the right of'.)

Shoemaker also considers heaviness (Shoemaker, 1994d p.28). As he says, what feels heavy to a child may not to an adult. Again, I share his intuition, that we pre-

theoretically experience heaviness as a non-relational property of objects, and not as what it is, a property defined in relation to our build and strength.

Given these mundane cases, it seems right to accept that the mere fact that Shoemaker's account requires us to experience, as non-relational, a property which is in fact a relation to an aspect of ourselves, is not in itself a valid objection to the view.

Which is not to say there may not be valid objections to it.

4.4.2 Why Does the Account Need *R**?

My own strongest reaction to Shoemaker's account is simply this: why can't our experience *simply* represent objects as being red? Why does it have to do so, in and by representing them as being *R**?

It is here, I think, that we find the most evident remnants of that approach to colour qualia which Shoemaker formerly shared (in large measure) with the Churchlands, Lewis and others (Section 2.2.4). For the functional role of colour experience is supposed to include the fact that it leads me to say "red feels *this* way to me", and when I say this, '*this*' is supposed to have some genuine referent. This much I accept, and wish to naturalise, myself (Chapter 5). What is further supposed (and *this*, I will question later) is that the only way to naturalize this subjective feel is to allow intrinsic (non-relational, non-functionally characterisable) aspects of experience (i.e. the 'qualia' of such earlier views) some mental role, in our account.

How can Shoemaker continue to allow such non-introspectible properties their (alleged) role in our mental lives? Shoemaker believes that, by allowing experience to represent relational properties of public objects, such as *R**, he can:

- Avoid requiring that intrinsic properties of experience are introspectible (which, he now accepts, they are not, on his own account of introspection)
- Avoid having an error theory of perception (as with projectivism, which is in other respects rather similar to his new account)
- Still allow for the possibility of mentally relevant, *intrinsic* qualia varying as between subjects who can make the same discriminations, and agree on a common language

That's why Shoemaker's account needs *R** – a property of the objects we observe, and one which varies in the way Shoemaker needs, whilst avoiding (or so he thinks) the problems which would arise from direct introspection of intrinsic qualia.

4.4.3 Why Does the Account Need Qualia?

Since qualia are no longer introspectible, on Shoemaker's account, one might also wonder whether his account actually needs *them*. Shoemaker's own response to this question is very brief; the bulk of the explanation seems to lie in this text:

> "This account needs qualia because it needs a way of typing experiences which not does consist in typing them by their representational contents. It needs this because only so can there be properties whose identity conditions are given by saying that things share a certain property of this type just in case they produce, or are apt to produce under certain conditions, experiences of a certain type." (Shoemaker, 1994d p.29)

As I understand it, the reasoning in the above goes something like this. Shoemaker's new account involves phenomenal properties, where these are relational properties of public objects: the property of producing an experience of a certain type. As such, if experiences themselves could only be 'typed' by their intentional contents, the account could not work: we would have no means to differentiate Jack's *Q1** from Jill's *Q2**. Shoemaker needs Jack and Jill's experience to represent two different, but equally true, facts about red objects. If we stick to typing experiences just by their intentional contents, and try to produce something like Shoemaker's account, we end up with circular proposals, such as the suggestion that my experience represents red things as having *R′*, where *R′* is the property of causing experiences which represent things as having *R′*. It is in order to break out of such circles – in order to have a way of saying *which* experience *R** tends to produce – that Shoemaker's account still needs qualia.

I think Shoemaker is right that he can avoid this particular circularity, in this way. But the resulting account is not without unresolved tensions, both as regards *what* is represented (Section 4.5), and as regards *what it is* for something to be represented (Section 4.6).

4.5 The Less Serious Acknowledged Problem

4.5.1 Which Relational Property is *R**?

Shoemaker's *R** is the property of producing, or being disposed to produce, quale *R*. But, as Shoemaker himself notes, this is only a rough definition. In more detail, several properties seem to qualify as candidates, with various pros and cons. For instance, *R** might be the property of occurrently causing *R* in a specific perceiver. Or of tending to cause *R*, in that perceiver. Or it might be the property of tending to cause *R* in a specific population of perceivers. Or, finally, *R** might be the property of occurrently producing

R in a subject related to it (i.e. to that property, *R**) in an appropriate way (Shoemaker, 1994d p.27).

If red (say) produces quale *Q1* in Jack, and *Q2* in Jill, then red has both *Q1** and *Q2**, on *all* of the above definitions. Initially, it might seem that the approach is spoilt for choice. But Shoemaker returns to the issue of deciding which of the above is the best candidate later in his paper (Shoemaker, 1994d pp.33-35), and there he admits that *none* of these candidates is ideal. Shoemaker considers four desiderata for phenomenal properties (i.e. properties such as his *R**: the relational property which colours are perceived as having, and the representing of which, in experience, gives experience its phenomenal quality, if Shoemaker's account is right):

1. That these properties should belong to external objects.
2. That these properties should be such that two subjects with experiences phenomenally the same should be seeing the same such property, and two subjects with experiences phenomenally different should be seeing a different such property.
3. Shoemaker also suggests that, ideally, such properties "should be ones that one can perceive something *not* to have by perceiving it to have an incompatible property of the same sort, in the way one can perceive something not to be red by perceiving it to be green" (p.34).
4. They should be properties which things can have when not perceived.

As Shoemaker notes, all of his candidates meet 1.; indeed, this has been his aim, throughout. Unfortunately, as he himself acknowledges, none of his proposed candidates meet all the remaining desiderata.

Properties defined with respect to *specific* subjects are not comparable between subjects, as 2. requires (i.e. they do not allow any sense to be made of the claim that my red looks like your red, say). Because of this, we would lose track of any sense of inverted spectra – it would not be possible to claim that my red is like your green, and your green is like my red. Properties definable in terms of *classes* of subjects still have this problem, for two creatures who are spectrally inverted would necessarily have different sensory constitutions. A creature with one constitution cannot perceive red to have to property of typically causing G in creatures with a *different* constitution. In Shoemaker's opinion, this rules out both the two candidate properties just outlined in this paragraph, since neither can be made compatible with the second desideratum, and since, as he puts it:

"The first two desiderata seem to me not negotiable." (Shoemaker, 1994d p.34)

This leads Shoemaker to prefer a property of the final type outlined at the start of this section; as he now puts it, the property of:

> "currently producing an R-experience in someone related to it in a certain way". (Shoemaker, 1994d p.34)

Indeed, he is happy to be somewhat more specific about *what* relation someone should have to the property, stating that his preferred candidate for *R** is:

"*is producing an R-experience in a viewer*" (Shoemaker, 1994d p.34, original emphasis)

Unfortunately, as he himself admits, this property, which looks like his best bet for sustaining the inverted spectrum intuition (2.), fails his two remaining desiderata (3. and 4.). Shoemaker notes that we could have achieved 3. by talking about the property of producing *R* experiences in me, but that would fail to satisfy 2.; and he looks at various other possibilities, but as he eventually says "unless I have overlooked something, there is no ideal candidate" (Shoemaker, 1994d p.34).

Shoemaker's explanation for this lack of an ideal candidate is to propose that whilst *R** is what we have just said (the property of producing an R-experience in a viewer), visual experience actually conflates two properties: the property of being *R**, and the property of being (plain, public, gerrymandered) red. His explanation for the lack of an ideal candidate is that some of the above desiderata apply to phenomenal properties (specifically, 1 and 2) and some to public colours (1, 3 and 4).

Of course, normal experience doesn't *seem* to have a two-fold character – we just see something as red. Therefore, as was noted above (4.3.2), Shoemaker considers this two-fold character of experience to be a 'cost' of accepting his view. As he points out, it requires us "to say that our experiences represent different properties that they do not distinguish" (Shoemaker, 1994d p.36). Shoemaker's defence of his position at this point (Shoemaker, 1994d p.36) consists, a) in noting that all philosophical positions have costs, and b) in suggesting that accepting this particular cost is not so different from accepting that experience actually represents relational properties, but doesn't seem to, even in some uncontroversial cases (which I have agreed is the case: 4.4.1).

I suggest that these lines of defence somewhat miss the point. How should we judge whether this is a significant cost, or not? Presumably this should hinge at least partly on what it is for an experience to represent one thing or another, in the first place. And it is

here, I think, that we come upon the most fundamental issue with Shoemaker's present view of qualia.

4.6 The Fundamental Problem

4.6.1 In Virtue of What Does Experience Represent *R**?

Shoemaker himself identifies – if only in a footnote – what seems to me to be the fundamental of tension in his account (he credits this line of objection to David Robb). Shoemaker asks:

> "In virtue of what does an experience having a quale represent an object as having a particular property, if phenomenal properties are what I say they are?" (Shoemaker, 1994d p.37, n.7)

The key point, here, comes in Shoemaker's own immediate response to this question:

> "It cannot do so in virtue of a causal relation between the experience and the property it represents – one cannot say that the causing of A by B is the (or a) cause of A" (Shoemaker, 1994d p.37, n.7)

Shoemaker doesn't say any more on the logic behind this objection, but it is worth looking at it more closely. Many fundamental issues come into play here. Are there *representational* features of experience at all? If so, in what sense? And if there are representational features, are they whole-system states, or sub-system states?

4.6.2 The Subsystem Story

Firstly we will look at the above issue, of causality in Shoemaker's account, with respect to a relatively standard account of representation in experience, in which certain subsystem states are reliably caused by certain external features, in normal circumstances. (Remember that, additionally, such subsystem features must play a certain role in the eventual behaviour of the agent, or else they are not the relevant representations – but this issue will not come into play until the next section.)

On such an account, state *P* (standing for physical representation, since *R* is already in use for something else), is reliably caused by, say, red things in the world. We expect there to be a detailed physical story about this, leading from the presence of items with a certain surface reflectivity or whatever (i.e. which are red), within a reasonably broad range of conditions, to instantiations of an internal state of this type.

But now consider Shoemaker's internal state, *R*, which is supposed to represent the occurrence of the property *R**. Remember, *R** is the property of 'tending to cause *R*'. If

we want to give an account of the above form, we should be looking for a detailed causal story of how some state arises, which reliably, in normal circumstances, indicates the property of 'tending to cause *R*'. To achieve this, we cannot just instantiate a detector for red (even though red does, *ex hypothesi*, tend to cause R in me), because we need the detector to be the kind of detector which would notice the difference, if red started causing *G* in me, instead. That is to say, we need a detector for 'causes *R* in me' as such, not just for something which in fact causes *R* in me. So far, no fundamental problem. I think we can imagine a system which locks onto external properties (such as red), *and to internal states* (such as *R*), and which can be adjusted to fire reliably, to those external states which tend to cause *R* in me. Now we reach the problem, for *R* is supposed to be the output of this system, as well as one of its inputs.

To give the normal kind of causal story, such a mechanism would need to detect red, *and* to detect *R*, and then (i.e. *because* of that) tend to produce *R*. That is *R* (or, at least, the causing or *R*), would need to cause *R*. Now, there are many different accounts of causality, but I think Shoemaker is right to concede that things don't cause themselves, in any reasonable sense, and nor does the causing of something cause that thing.

The issue perhaps comes more clearly into focus when we try to imagine what the internal causal story is supposed to be, for a system such as that just discussed, which is meant to produce *R* partly because it is producing *R*. No such story is available: of course something can produce *R when* it is producing *R* (everything which produces *R* does so); but not *because* – there isn't the independence of characterisation of cause vs. effect needed to make out a causal story, on any plausible account.

By the way, I am certainly not claiming that it would be difficult to come up with a physical system which more or less reliably produces state *x*, when it encounters states which tend to produce state *x* in it. If that was all we were trying to engineer, things would be too easy, for every physical system more or less reliably produces state *x* when it encounters states which tend to produce state *x* in it. The problem is that a *causal* story as to why (how?) states represent what they do could never justify the claim that such a state, *x*, represents 'tends to cause *x*'.

Perhaps we should just accept the limitation Shoemaker identifies, on the kind of account which we can possibly give, of representation of these 'phenomenal properties'. As Shoemaker's own brief comments correctly imply, causal accounts are not the only accounts on offer in the literature, of the representational status of internal states. So perhaps (if his theory is right) we should just accept this as a constraint on future

theorizing about the nature of representation in the most general case. However, there still seems to be something problematical here. Can it really be right that we *can't give* a causal story of how some state comes and goes, which represents *R**? Once again, we seem to have a rather strong threat to physical explicability. Depending on one's intuitions, this may or may not seem worrying. At which point, it might seem hard to know what to say next.

I think part of the problem with pinning down what is fundamentally at issue, is that we are currently considering accounts featuring *subsystem* representational states. I am no longer convinced that such accounts can work as advertised, and Shoemaker, too, has said that he is *not* committed to the presence of such separable subsystem states (Shoemaker, 1990 p.67). Considered in the context of Shoemaker's present position on qualia, as just outlined, this might sound strange. But remember, Shoemaker's *R* is just *some* intrinsic (and representing) property of the agent. Arguably, it could be a whole system property. All that it *has* to be, for Shoemaker's account to work, is an intrinsic property (i.e. *something* about the agent which is not capturable by a purely functional description of the agent's mental states).

Perhaps, then, we've missed the point? Perhaps considering whole system states will relieve some of the apparent tension in the account? Unfortunately, the situation is quite the opposite. Remember that Shoemaker (just like all other authors whose positions were critiqued in Chapter 2) is arguing for a continued role for intrinsic properties in the characterisation of experience, precisely because he believes such properties are required, to *complete* the naturalisation of mental states considered in their public, *a priori*, functional role. Certainly the one thing Shoemaker doesn't want is an account *incompatible* with mental states, considered thus. Unfortunately, I will argue, this is what he has.

4.6.3 The Whole System Story

The problem for Shoemaker is that functional accounts of the personal level *are* causal accounts. For instance, a typical causal, functional story from perception to action might go as follows: certain physical states of affairs (combined with certain physical and mental states of the subject: eyes open, attending) cause certain mental states of affairs (perceiving), which interact with other mental states of affairs (believing, desiring) to produce eventual action. This is how functional accounts work. Shoemaker has not offered, or intended to offer, anything different.

An experience[91] of red, on a perfectly normal, functional account, is that state which is normally caused by red things, and which normally causes red appropriate behaviour (in interaction with other mental states, including motivational states). For more detail on the nature of the specific causal relationships which must be in play for there to be *bona fide* perceptual experience, see the Appendix, where I review Noë's highly plausible recent treatment (Noë, 2003) of this topic. Some such account (say Noë's, if it is the right one) is the *a priori* nature of the experience of red (in a relevant sense, no doubt narrower than common usage). There has to be *some* such analysis, according to *any* functional account; and it is a functional account which Lewis, the Churchlands, and Shoemaker all mean to endorse.

Now, in the previous section I discussed the point which Shoemaker certainly meant to acknowledge (though it already looks worrying): that his analysis of the relation between *R* and *R** precludes a causal analysis of how *R* represents *R** – if *R* is treated as a sub-system property. The problem is that the same line of reasoning applies to experience understood as a whole-agent state, with *R* understood as some intrinsic feature of that state (but where one now remains neutral as to whether *R* is a subsystem or whole system intrinsic property).

Consider Shoemaker's account of the relation between *R* and *R** once again, but now at the whole system level. The whole system state, of experience of *R**, is meant to (normally) occur when *R** occurs. Remember, *R* is one of the experience's *defining* features (see Section 4.4.3): it is *that* type of experience – the type which has *R* as a feature – which is meant to typically come and go as instances of *R** in the environment come and go, if Shoemaker's account is right.

Now, once again, a physical account of the coming and going of such a state, which is supposed to be sensitive to *R** as such (i.e. not merely to something which is in fact an instance of *R** – see the previous section), must include physical sensitivity to *R*. In any detailed causal account of such a process, we'd be trying to understand *how* a certain physical state arises, in response to another state. Once again, it is not possible to give a causal account of how an experience with intrinsic feature *R*, representing *R**, arises (partly) in response to its own existence.

[91] Where 'experience' is to be read success-neutrally, i.e. seeing or seeming to see (c.f. Sections 2.3.3 and 5.5).

As before (in the argument at the subpersonal level, of the previous section), of course an experience *can* arise, when it exists: everything 'arises' when it comes to exist. The problem is that it can't be said to arise (partly) *because* it exists.

The issue, then, is not just what Shoemaker acknowledges in a footnote, that he has ruled out the possibility of subpersonal explanation of a very standard type – though he has, and that might be issue enough. Worse yet is that the nature of the relation between *R* and *R** also rules out a causal relation between the experience and what it represents, *purely at the personal level*: an experience with intrinsic feature *R* can't be said to exist partly because it has intrinsic feature *R*, if the 'because' is supposed to have the force of 'in response to the presence of'.

It is especially important to note that Shoemaker's account of introspection of *intentionally defined* states, which I have endorsed, *does* admit of a causal analysis in the sense at issue here. The state of believing that I see a red ball can (and does) arise because I see a red ball (and because I turn my attention to this relation, between myself and the world, and so on). The force of the causal analysis here is not the deterministic, objective kind of physical cause which is felt by some to be fundamental[92], but rather something more like the manipulability notion of cause (Woodward, 2001/2008), wherein it would be rational to attempt to affect C, in order to affect E. Something like this causal analysis can certainly be applied to perceptual experience (on Noë's account of it, say), and to introspection of rationally defined states. But it is ruled out when we try to understand what it is for Shoemaker's *R** to be present to a subject even (if he is right) in the most basic case of experiencing colour.

As such, Shoemaker's account of qualia doesn't just require some clever footwork, as regards the relation between intrinsic states and what they represent. Instead, it throws into doubt the entire causal account of the nature of the mental which it means to complete. Suddenly there are experiences wherein being an experience of *x* does not inhere in any form of the causal, *a priori* functional relation between the state of experience and the experienced state which this account of qualia was supposed to be trying to support and complete.

[92] Though felt by others (including the present author) to be chimerical (Price and Corry, 2007).

This is all clearly very much related to the problems identified in Chapter 2, where it was argued that neither phenomenal concepts[93], nor knowledge possessed by exercising them, could ever be naturalised on any normal scientific account either. The difference in the present case – and perhaps it makes the problem even more stark – is that the property of which we have 'inexplicable' knowledge, on Shoemaker's account, is not simply a property of the subject (as it was in the case of the problematic kind of phenomenal knowledge discussed in Chapter 2) but is instead presented as being a *bona fide* relational property of public objects. Even so, knowledge of these phenomenal properties[94] cannot be explained in the normal way, and *cannot be accounted for functionally*.

4.6.4 Knowing Qualia

If Shoemaker's new account of the nature of qualia is right, then in introspection we *can* still attain exactly the kind of 'more specific' knowledge of what is going on is us, via introspection, which I discussed at the start of this chapter. This is knowledge which is more specific than simply 'some physical state compatible with the functional state I am in, but I know not which'; and it is knowledge which, on Shoemaker's own account, we cannot gain *directly* via introspection. Instead, as Shoemaker himself clarifies, on his present account we would gain the knowledge of which intrinsic (qualitative) state we are in via introspection, supplemented by an understanding of the general theoretical concept 'quale':

> "Introspective awareness is awareness that. One is introspectively aware that one has an experience with a certain representational content, and with the phenomenal character this involves. And if one reflects on the matter, *and has the concept of a quale*, this brings with it the awareness that one's experience has the qualia necessary to bestow that content and that character." (Shoemaker, 1994d p.28, emphasis added)

Perhaps this sounds a little mysterious – what, exactly, is supposed to be involved in this process whereby one introspects and "reflects on the matter" (Shoemaker, 1994d p.28)? How do we thereby come to know which intrinsic (hence, on Shoemaker's own account, non-introspectible) property is instantiated in us? In the end, though, I think we

---

[93] On the phenomenal concept strategists' analysis of them (though not necessarily on some more moderate analysis).

[94] On Shoemaker's analysis of them (but again, perhaps not on some more moderate analysis of them; see Section 5.3.4).

should accept that Shoemaker has successfully pushed all the problems to where he thinks he has pushed them. *If* we can once be persuaded that experience *already* (even in the non-theoretically informed) represents *R**, then the next steps (the introspecting, the 'reflecting on the matter') can probably be allowed to go through.

It is by the use of *R** that Shoemaker has managed to lever intrinsic properties back into experience (of course they never left, from his point of view). He believes he must keep a place for intrinsic properties, in order to account for the inverted spectrum; and indeed he must, if we are to have behaviourally undetectable inverted spectra. But the price of all this seems much too high. The central issue is not to do with the two-fold (4.3.2, 4.4.3, 4.5) or relational (4.4.1) character of the contents of experience. Instead, we have two inter-related, fundamental problems, both arising from an issue which Shoemaker explicitly (if very briefly) accepts, when presenting his account.

Firstly, his analysis entirely rules out a very standard kind of causal account, as regards how subpersonal level physical events instantiate the mental, and such as we might otherwise reasonably hope to see, in some form or another, in a viable naturalistic account (4.6.2). This already seems like too high a price to pay, to me, but it evidently doesn't seem so, to Shoemaker. So be it.

The second of the two inter-related issues comes into view when we look at how this same restriction on the nature of the 'representational relation' applies at the personal level. In this case, there looks to be a problem even within the terms of reference of Shoemaker's own approach. For Shoemaker's account is presented as a way of naturalising our functional understanding of the mind, yet it makes certain fundamental *mental level* relations (never mind any subpersonal relations which may instantiate them) non-causal, hence non-functional. The upshot of this is that the *a priori* (in Lewis' sense) nature of mental *relations* is *not* fully functional, on this account, even though the account was presented as a *defence* of such an analysis. In all honesty (and though this it is, at best, quite deeply implicit in his work) it may be most accurate to say that Shoemaker has always been aware of this cost, too, and willing to pay it (c.f. Shoemaker, 1975 Section 6). But these costs are very high. Do we really need to pay them?

Shoemaker's own account of introspection *appeared* to be telling us that you just can't gain knowledge of intrinsic properties, by introspection (not directly, not indirectly). As far as I can make out, it still does tell us that, as long as we accept that the *a priori* nature of the mental is causal; for then what we see is only what there is (it

cannot include relations to intrinsic properties), and what else we can discover, in introspection, is 'only' our (at least in principle, publicly verifiable) relation to what there is.

So, what happens if we're not prepared to pay the costs which Shoemaker is prepared to pay? What happens if we keep a fully causal account of the mental level, and accept what Shoemaker's account of introspection then tells us: that you just can't gain knowledge of intrinsic properties by introspection. Is there any hope of retaining qualia? In the next chapter, I will argue that there is.

5. *A Space of Reasons Analysis of Qualia*

5.1 Introduction

I have argued in Chapter 3 that it is in the nature of introspection to give us access only to properties of a space of reasons as such; that is, to analytically public mental properties (such as states of belief, desire, perception, etc.) and not to any intrinsic properties of these states. That is to say, we do not have access to physical role fillers, nor do we have access to intrinsically mental role fillers (pure phenomenal properties); nor yet do we have access to physical role fillers under some fundamentally private mental mode of presentation (as the phenomenal concept defence of physicalism would have it – see Section 2.2.5).

I have also argued, in Chapter 2, that qualia must be introspectible, at least on occasion, at least in us, who seek to explain them. Combining these two arguments rules out many analyses of qualia, including many which brand themselves as physicalist (various such accounts were presented and critiqued in Chapter 2).

I argued in the previous chapter that these considerations even rule out Shoemaker's own present account of qualia, albeit that it is Shoemaker's account of introspection on which I am drawing.

Perhaps the most 'natural' approach to such difficulties is to adopt Dennett's claim, that there simply *are* no qualia, in anything like the sense we naïvely suppose (Dennett, 1988; Dennett, 1991). Indeed, I have already clarified that I fully agree with Dennett that the 'qualia' of the 'strong phenomenal realist' approaches canvassed in Chapter 2 cannot exist. Nevertheless, I tried to leave open in that chapter the possibility of a moderate phenomenal realism, in which we seek for qualia amongst the properties introspectible on some independently plausible account of introspection.

Using the account of introspection argued for in Chapter 3, this means that qualia (if we are to find them at all) must exist within the properties of a space of reasons, as such. In this chapter, I will present an analysis of qualia which identifies them as just such properties. I examine two of the most standard examples of qualitative feels – colour qualia, and pains – and argue that it is possible to find a place for these in the

structure of a space of reasons as such (i.e. without any mention of how such a structured locus of action is physically instantiated – although of course it must *be* physically instantiated; and equally, without any mention of any intrinsic, contingent, non-publicly accessible mental properties). If true, this makes qualia both necessarily mental (not merely contingent role fillers), and introspectible, which seem to me highly desirable features of the account.

To set the scene for the analysis of qualia to be offered, I will first of all add to the characterisation of the space of reasons account of the mental already given in Section 2.3, by making some observations about the ineliminable role of *affect* in the characterisation of action for reasons.

5.2 Affect as Modification of a Space of Reasons

My emphasis on rationality, and on the conceptual, should not be taken to imply that I am dealing with some kind of cold rationality, divorced from an engaged stance in the world (an idea which is anyway, as McDowell says, only a "dubiously intelligible kind of thing": McDowell, 1994 p.117). However, I have already indicated (Section 2.3.2.1, Section 2.2.7, etc.), that affect and motivation would come to play a central role in the analysis of qualia offered in this thesis. What we need to note now, is that as long as we think of rationality in 'ability' terms (c.f. Evans, 1982 p.101) (i.e. as inhering in at-least-counterfactual behaviour of the appropriate type), then we cannot conceive of an agent as having a space of reasons for action, *without at least implicit acknowledgement of affect*.

For no mere collection of facts (including indexical and demonstrative facts) is sufficient to lead to any rational act, whatsoever. Hume makes the same point in saying that "Reason is, and ought only to be the slave of the passions, and can never pretend to any other office than to serve and obey them" (Hume, 1739-1740/2000 §2.3.3.4; as quoted in Froese, 2009). To put the point crudely: I am about to be crushed; infants are about to be killed; without some affect, somewhere, so what? *Perhaps* it may be that there are ethical or other normative facts which can be fully expressed as conceptually articulated premises ("killing is bad"; "peace is good"). But it is certainly not the case that most people, even when acting rationally, act on the basis of such explicitly

formulated premises. Instead, an agent *just* loves its offspring, or loves freedom, or the search for knowledge – or, more prosaically, is *just* hungry, tired, etc[95].

It seems that the right way to put affect into a space of reasons is instead to say that food just is desirable, for instance (at least, when an agent is hungry). Indeed, it might be better to turn things about and say that a certain way of acting is what hunger (again, as an example) is. Hunger is not some internal rule to be followed; hunger is not even a reason (though *food* might well be); hunger is a certain structure of action for reasons.

Now, most would surely agree that however hunger works, it does not involve an agent being *aware of* a rule. The point I am making is not just this platitude. My point is that it would miscapture the space of reasons of an agent to say that being hungry involves the agent *following* a rule, in *any* sense. To see why, let's first briefly discuss what it is to follow a rule, in the most basic case.

We get caught in vicious circles if we say that, for an agent to follow a rule, they have to be *aware that* they are following a rule (c.f. footnote 95). What I think we must say, for a rational agent to be said to be non-metaphorically, and *qua* rational agent, *following a rule*, is that the agent is at least aware *of the rule* as something-to-be-followed. Such an agent can be said to be 'acting in accordance with *that*', because the agent knows what it is (in a practical sense) to 'act in accordance' with something, and is aware of '*that*' as something being acted-in-accordance-with (again, in a practical sense).

This is simpler (more basic, less demanding of an agent) than the rule-following case we might most naturally think of, in which an agent is aware of a rule *as a rule*. This simpler case is, I think, the most basic case in which it can be correct to characterise an agent's space of reasons by saying that the agent is following a rule. But even this is more high-blown than much of our practical rationality. Some of our actions for reasons are like that – they involve following rules in that way. But not all.

Indeed, to return to the example at hand, we cannot analyse the basic case of being hungry like that: we won't find something which an agent is 'acting in accordance with', in the above sense, when an agent does what it is rational to do when hungry.

My own proposal is that instead we should think of hunger (our initial example) in terms of what might best be called a 'modification of a space of reasons'[96]. At certain

[95] Indeed, I am far from the first to point out that whatever rules we might follow, it *cannot be* that all our acts are rationalized by explicit rules (Carroll, 1895; Wittgenstein, 1953/2001).

times in an agent's life it will simply be desirable to seek out food. That is, food on its own becomes a reason for action; there is no further reason; the food itself is 'presented desirably'.

Why am I describing seeking food as a modification of a *space of reasons*? What has rationality got to do with this at all? My proposal is that hunger (*qua* mental state, rather than *qua* state only metaphorically ascribed to some very simple agent, c.f. Section 5.4.3) exists exactly when a physical locus of action for reasons (that is, a mind) undergoes such a modification. For there to be a mind at all, we need there to be an appropriate complex rational structure. There need to be lots of rational categories. The categories (or rather, aspects of the world as under those categories) need to be co-relevant (relevant at the same time) to the agent, in its choice of action. There need to be lots of rational transitions present, between states which are defined in terms of these co-relevant categories[97].

Then, for there to be hunger (as I am here analysing it) we need to see this space of rational action modified in such a way that food (public objects, which are food objects for that agent) itself becomes, for that agent, a reason for certain actions.

It may be alleged that there is a kind of circularity in this proposal, but I do not believe that it is viciously circular. It is true that you cannot identify the rational categories which a creature is thinking in terms of, without identifying the creature's motivational structure; and that you cannot identify the motivational structure without identifying the rational categories. However, such issues as there are here are (pretty clearly) direct descendents of the issues involved in identifying the beliefs and desires of an agent. And such issues as there are, are soluble (Lewis, 1970): it is a non-vacuous claim, subject to empirical verification, that an agent has a given belief-desire (or perhaps one should say category-affect) structure[98]. To make such claims is to say verifiable things about the at-least-counterfactual *action* structure of the creature.

---

[96] In a manner which intentionally echoes the adverbialist notion of 'modification of a subject' (for more on the relation between this account and adverbialism, see Section 5.5).

[97] All of this is meant to suggest a conceptual account of mind, and the applicability of Evans' *Generality Constraint* (Evans, 1982) to these concepts. For a little more on these issues, see the Appendix.

[98] I am attempting to identify states which are of the same general type as belief and desire, but more basic than them; to wit, perception and affect (with affect construed as just described). It turns out that this parallels a current move in animal ethology, which claims that entry-level mind-reading abilities in other animals require only that the animal is responsive to states of the other at the level of a *perception-*

So, we cannot get a space of reasons for *action* until we have motivational structure (for all that this is often left implicit). This means that an observer can have no reason *whatsoever* to say that an agent (*qua* locus of practical action for reasons) sees red, or green, or a dog, or a house, unless the observer identifies not just some categorial structure but, at least implicitly, enough of the agent's affective structure to make it the case that the categorially structured states are states where the agent has reason to do something, rather than nothing at all[99].

5.3 Colour Qualia

5.3.1 Necessarily Mental Qualia

How does all this relate to qualia? It would be slightly too simple to say that I'm equating qualia with these affective states, although it would be close. Here is the proposal:

> The quale associated with a given perceptual category[100] is identical to the sum total of the subjective effects of the related objective property on a given subject, at the space of reasons level of description.

In order to clarify what I mean by subjective, here, I will briefly recall some points from Chapter 2; for I am not going back on my claim that our notion of the mental is co-extensive with our notion of the publicly observable, at least counterfactually behavioural, mental level; I certainly don't mean to reinstate private mental objects, in any of the senses which Wittgenstein, for instance, argued against (Wittgenstein, 1953/2001).

A crucial observation, in order to clarify what I do mean, is to reiterate that you have not fully specified a state, as a state of a space of reasons, when all you've specified is

---

goal psychology, rather than a full-blown *belief-desire* psychology (Tomasello, Call and Hare, 2003a; Tomasello, Call and Hare, 2003b; Call and Tomasello, 2008).

[99] In allowing for such a thing as *perception-goal* psychology, Call and Tomasello's recent work (see preceding footnote) can at least arguably be read as endorsing the claim that it's not just mind-*reading* which comes in at the perception-goal level, it's *mind*. There need not be the full-blown structure of explicit belief-as-such behaviour, in order for there to be action in the space of reasons (c.f. Hurley, 2003).

[100] For instance, the perceptual category which picks out the public surface property red (where red is construed as a gerrymandered public property, and *not* as the property of causing a certain quale in a observer).

the categorial structure of that state and the particular 'propositional attitude' relation (seeing, remembering, imagining, etc.) which the agent has to (or as if to) the world under those categories. Thus, to say that an agent perceives a red flower on a green field, is so far to say nothing at all about which action the agent will take. But a space of reasons for action can only have been fully characterised when you can say what is reason for what – when you can link perception to *action*.

Therefore, it seems clear enough that two agents could agree exactly on what there was, and could even agree exactly on the language for what there was, and yet could *act* entirely differently, when faced with the same situation (Section 2.2.7). This is because you can specify everything about the categorial structure within which an agent perceives, and everything about the words which an agent uses to label what it perceives, without having said anything about what it will do, when it perceives some given thing. That is still left to specify, by specifying the relevant affective structure; therefore we can coherently propose two agents who differ only in this structure.

Moreover (and this is the reason why it would be too simple to say that I'm identifying qualia with affective states), it seems that there is at least one further way in which the two agents above could differ. Nothing about what an agent perceives, or about what words the agent uses to describe what there is, determines the agent's *associations* between perceived categories of things, either. Thus, one agent may be reminded by red of blood and death, the other of celebration and success (or, if they are Kant, of "heavy cinnabar" – Kant, 1996 (1781/1787) A101).

All of this – the associations, the affect – is free to vary, as between two agents who agree exactly on which things are red. And all of these facts are indeed subjective facts in the sense of Section 2.2.7. They are not *private* states (describable only in a 'language' not communicable to another), but they are facts which go beyond *what* is perceived, and which go beyond the bare essentials necessary in order to specify that a subject has a mental relation (perceiving) to what is perceived. We move from merely specifying *what* the subject sees and *that* they see it[101], to specifying aspects of what they are going (at least *ceteris paribus*) to do about it.

[101] That is, we move beyond the basics required to establish *that* a subject has a perceptual relation to something (which is already enough, on this account, to establish that some subjective facts obtain, but not yet to say which ones).

It is my contention, then, that this – the subjective effect that red has on me, in the above sense – is what I mean, when I refer to the subjective phenomenal quality of my experience of the shared, public property red.

5.3.2 Introspectible Qualia

It is crucial to note that, on the account of introspection argued for in Chapter 3, these qualia are introspectible. To see why, we can introduce the concept '[the way red affects me when I am having a perceptual experience as of red]'[102] ('affects me' here means specifically, affects my mind, *qua* locus of practical rationality). Then, once again, we can run Shoemaker's style of argument (Section 3.7): if the running of my rational mind is indeed being thus affected, and if I turn to attention to whether or not it is, and if I conclude wrongly, then this is a failure of rationality.

Once again, this is not to say that any given agent *must* be able to take this rational step – but it is to say it would be no more than 'mere' rationality, to be able to do so. All of this follows given two elements which have been argued for already: that Shoemaker's typical line of argument can be adapted to *any* property of a space of reasons as such (argued for in Section 3.7), and that the above property *is* a property of a space of reasons as such (just argued for, in Section 5.2).

5.3.3 What Mechanism?

It might be thought that it remains unclear what I actually mean, in saying that this step can be achieved by 'mere rationality'. Specifically, what, exactly is supposed to be physically involved in instantiating such rationality in an agent? This thesis does not try to answer this question in any detail, so I raise this point only in order to recall what I have already tried to establish (Sections 3.5.5 and 3.6.4.3): that this kind of introspection is, at least, fully *compatible* with physical implementation, in a real agent. The present analysis of qualia is compatible with what the Shoemaker-Sellars theory of introspection *allows*: that there need be nothing magical about a creature which *could*

[102] Just as an agent need not have some *other* notion of 'giraffe', say, in order to possess the concept 'giraffe', so the agent need not have the concepts which I have used in order to describe what is picked out here, in order to be sensitive to it. This is why I have used the square bracket notation. They do, though, need to be sensitive to what I have picked out, as such, when it occurs (at least on favourable occasions, etc.).

make such a transition; such a transition *could* occur in a physically explicable way (c.f. Section 3.5).

It is of note that this positive result does not hold for Shoemaker's own current account of qualia. Features of his account of qualia mean that it cannot inherit this attractive feature of his account of introspection more generally (and, indeed, that is my central objection to his current account of qualia, Section 4.6).

5.3.4 *R** Again

Nevertheless, I said that aspects of Shoemaker's account would remain recognisable in my present account, and here I'll explain why. Exercise of the above proposed concept '[the way red affects me when I am having a perceptual experience at least as of red]', requires that the subject be sensitive to something which only occurs when they are having an experience at least as of red.

It should be emphasized that simply having an experience of (public) red is less demanding than this; it does *not* require exercise of the above mentioned theoretical (or at least folk theoretical) concept. Having an experience of red is no more or less than the bringing of (public) red into a space of reasons (or, as it may be in the case of illusion etc., *acting* as if red were within a space of reasons, when it is not).

This simpler state, the state of experiencing red, can certainly exist in a creature without the conceptual sophistication required to entertain the more complex state (though the converse is not true). For this reason, a creature can *have* qualia, *in the very same sense in which we mean it of ourselves*, even if that creature cannot *think that* it has qualia.

What, then, of Shoemaker's *R** (or what is certainly a recognisable descendent of it): the relational property which red has, of affecting me that way? Red certainly *has* such a property. Not only that, I can certainly at least *infer* that it does, given that I am already aware of red, and (introspectively) aware of red's effects (from now on, call red's effects on me q_r, for 'quale of red', for short). In that case *R** (or this account's descendent of it) is the property red has, of causing q_r in me. Given that I'm sensitive to red and sensitive to q_r then, it seems, I can become sensitive to red's having property *R**.

However, becoming perceptually aware of red is a noninferential transition (it is a basic mental act – there is no further *mental* explanation to be had). Equally, my becoming aware of the property q_r (the property of a space of reasons which I'm

proposing be identified as the quale of red, in me) is *also* a noninferential transition, on the account of introspection I'm endorsing. Again it's something I just can do[103].

The further point of this section is to observe that there seems no logical bar to my learning (c.f. Section 3.5) the ability to become *noninferentially* aware of red *as having the property of affecting me in the way in which it does*. That is to say, if I attend to something red, and then attend to the question of whether or not it has *R** (the property of affecting me[104] in the way in which red affects me), then it will be a failure of rationality if I conclude other than that it does.

This process, of becoming noninferentially aware of red as having the property *R**, seems to me neither fully perception, nor fully introspection; it essentially involves both. But there seems no bar to my (or some possible agent's) learning to do it.

I realise there is a possibility of misunderstanding here, so I should say that I am emphatically *not* claiming that we see red in and by seeing it as having *R** (as Shoemaker claimed). My account does *not* inherit that feature. First and foremost, in the most basic case, we see red – we are sensitive to red things. It is, surely, undeniable that red *does* have a certain (direct or indirect; major or subtle) subjective effect on us – as it must if we are to have any reasons for *action*. But we need neither be *aware of* this effect, nor of red as having it, in order to act.

5.3.5 Awareness of q_r

What, then, is involved in being aware of the effect itself (i.e. of q_r; we will return to *awareness* of *R** shortly)? As far as I can make out, there is no sense to be made of being aware of q_r (thus, of qualia, in general), except in a subject who is *at least somewhat theoretically informed*: who has an idea of what it is for some public property to have some subjective (not intrinsically private, but subjective) effect on them.

However, as clarified at the end of the preceding section, a subject does not have to be *aware of* their qualia in order to *have* them, and have them in the same sense in which we mean it of ourselves, on those occasions when we are aware of the fact that we have them.

Having said that a subject needs to be somewhat theoretically informed, to be aware of their qualia, I should clarify that I do not think the subject's theory has to be exactly

[103] Or, it might be better to say, there is no physical bar to my, or some agent's, being able to just do this (with physical, but no further *mental*, explanation in the offing).

[104] 'Me' *qua* rational subject.

correct. They might, for instance think of the effect which red has on them as a fundamentally intrinsic property, rather than as the public, at-least-counterfactual-behavioural property which I am claiming it is. But they have to have *some* idea of 'the effect red has on me'.

5.3.6 Awareness of *R**

The requirements for awareness of *R** follow similarly: if one has a theoretical notion *something* like secondary quality (i.e. the property of having a certain subjective effect on oneself) then, I am proposing, one can hardly escape knowing, directly and noninferentially, that colours have such a property, if and when one turns one's attention to the matter. I *can* (and, in the most basic case, do) think of blue as 'that property', out there (i.e. public, gerrymandered blue). But I can *additionally* think of the effect blue has on me (q_b). And therefore, I can *also* think of blue as the property which has that effect (*B**).

It is quite possible that in the experience *of the theoretically informed* all of these properties are available, and perhaps even conflated, since they all co-occur. Indeed, I suspect that acknowledging this complex situation may be one crucial step in helping to resolve our perplexity over qualia.

5.3.7 Some Clarifications

5.3.7.1 These Qualia Do Not Represent

Note what these qualia are not. I've already (5.3.4) said that on the present account we *don't* see red by seeing *R**, nor by being aware of q_r.

Of course, all this means is that *R** and q_r are not representations in a very trivial sense: they are not in view as representing properties, for the subject. But, as far as I can see, there is very little about the qualia of this account which makes them representational properties at all.

This is because the qualia of this account are not accessible to thought at all, *independent of a thinker's having a grasp on external, public properties*. There are no inner states of which we can be more certain than we can of the world (c.f. Martin, 2006). Similarly, these qualia most emphatically are *not* the highest common factor which the direct realist is keen to deny (see Section 5.5). To the extent that 'representation' survives in this account at all, we have an experience which

'represents' red, when and only when we *either* have public red in our space of reasons[105], *or* are responding as if we did, when we don't.

Therefore, it is unclear whether there is any remaining good motivation to call perceptual states on this account representations except, perhaps, for highly misleading historical precedent. For nothing which represents in the mundane sense need be in view, for either the subject or the theorist, in order for such a state to be fully in view as what it is: an introspectible state of a space of reasons. The fact that this point applies equally from the theorist's and the subject's point of view should not be surprising since, on Sellars' account, each uses the same mental concepts with the same conditions of application.

5.3.7.2 These Qualia Do Not Require Introspection

I have emphasized in 5.3.4 (and elsewhere) that these qualia can be possessed, in the sense which we mean it of ourselves, even by creatures which cannot introspect.

It is also worth mentioning that, by the same token, the present account is not a HOT account (e.g. Rosenthal, 1986; Dienes, 2004). For example, it is in no way my wish to claim that creatures with no concept of their own qualia 'almost' have conscious minds, but not quite, for lack of even the potential for those higher-order thoughts which are required to render the lower order thoughts conscious (at least on the most standard form of HOT account).

Nevertheless, the present account is consistent with the intuition behind at least some versions of HOT: that it is of the nature of phenomenally conscious states to be available to introspection (this is further argued for in Section 6.2). This follows directly from features already present in Shoemaker's account of introspection, as it applies to properties of a space of reasons as such, when combined with my additional claim that qualia are, indeed, such properties and are no kind of contingent role filler.

5.3.7.3 These Qualia Do Not Involve Confabulation

I should also emphasize that these qualia are *not* confabulatory. Put another way, when we introspect our own qualia successfully (as we can and do[106]) then what we take to

---

[105] This should be strengthened to 'have (public) red in our space of reasons in the right way'; the account of 'the right way' which I would endorse is Noë's account (2003), reviewed in the Appendix.

[106] As we would *always* do, when we set our minds to it, *if* we were ideally rational (which, of course, we are not and no real agent can be).

'be there' is really there, and is as we take it to be: there really is a subjective way red affects me, and I really can form a noninferential, introspective-demonstrative concept of which way that is. This is at least a part of what I mean when I say that this account amounts to a moderate phenomenal realism (Chapter 2).

This too is a feature inherited from the Shoemaker-Sellars account of introspection, for properties of a space of reasons as such, when combined with my claim that qualia are such properties. For introspection on this account is not an intrinsically confabulatory exercise; certainly, no more nor less than perception is. Just as perception, when successful, carves up the world into categories which allow a rational way to live around here (this is my terminology, not Shoemaker's or Sellars'), so too introspection carves up the rational mind itself into ways which are rational ways to respond to rational minds around here. And in both cases, one can endorse McDowell's direct realist manifesto: "when one is not misled, one takes in how things are" (McDowell, 1994 p.9)[107].

Of course, I am not saying that *every* introspective act hits its mark. I am not even saying that there are *any* introspective acts, wherein we can be certain that *that very act* hits its mark. But I am suggesting that, for introspection as for perception, it is not coherent to suppose that we are subjects who are in an epistemic situation such that no such acts hit their mark (c.f. Martin, 2006 for more on the direct realist response to various forms of scepticism).

5.3.8 A Plausible Candidate?

If I'm right, then these qualia are introspectible; and they are 'the subjective effect which red (say) has on me' (in the sense of Section 2.2.7). I have already argued (Section 2.2) that if such properties could be found, they would be suitable candidates for identification as 'qualia', in virtue of providing a plausible naturalisation of the inverted spectrum intuition. The additional contribution of this chapter, of course, has been to argue that such properties can indeed be found, by means of the above analysis.

---

[107] However, I am well aware that there is strong evidence that we often *do* confabulate, when self-attributing mental states (Nisbett and Wilson, 1977; Lillard, 1998). Briefly, I would argue that this shows that our rationality in introspection is much less than ideal, and/or that much of what we *take* to be introspection is not really introspection at all, but is inference from third-person evidence. I don't think any of this is sufficient to rule out the claim that *bona fide* introspection should be analysed on the Shoemaker-Sellars model, although of course much more could be said here, on both sides of the debate.

I also suggested, in Chapter 2, that such properties might hold out the hope of a plausible naturalisation of various other problematic intuitions about the nature of qualia. I will return to that issue in Chapter 6.

Before doing so, however, I will argue that the account just presented for colour qualia can be extended to the problematic case of pain.

5.4 Pain

5.4.1 Pain on Shoemaker's Account

Shoemaker does argue that his new account of qualia (Chapter 4) extends to pains. As in the case of Shoemaker's comments about pain, made in the context of his extended analysis of introspection (see Section 3.3.6), the comments on pain made in the context of his most recent analysis of qualia are all too brief.

Just as a colour experience, on Shoemaker's present account of qualia, represents a coloured surface as having a certain, relational "phenomenal property", such as *R**, so also, he suggests, pain experience represents a body part "as having a certain phenomenal property, namely hurting" (Shoemaker, 1994d p.31). That is, he suggests that the account extends, and that, in the case of pain, we happen to have an ordinary language word for the property which body parts are represented as having, in pain experiences.

He also proposes the same kind of account as for colour qualia, of our theoretically informed access to the quale of a particular pain:

> "Going with this perceptual awareness of the foot hurting is introspective awareness that one is having an experience of one's foot hurting. And this should not be thought of as an inspection by inner sense of the quale which gives the experience this introspective character. There is no such inspection. The kinds of awareness there are here are, first, perceptual awareness of the foot, second, introspective awareness (which is awareness that) to the effect that one is having an experience which if veridical constitutes such a perceptual awareness, and, third, the theoretically informed awareness that the experience has qualia which enable it to have the representational content it has." (Shoemaker, 1994d pp.31-32)

Again, I outline this account not in order to fully endorse it, but in order to comment on it, and to contrast it with the account which I will offer.

5.4.2 Problems With Shoemaker's Account

Once again, we see that Shoemaker's qualia are (on his new account) *not* directly introspectible, as they cannot be, if they are intrinsic properties (role fillers, rather than essentially mental roles). Equally, once again, one's experience in Shoemaker's most basic case, represents one's foot as having a property, *where one need have no idea of what it is, to have this property*, in order to have such an experience.

However, if the 'space of reasons' account of perceptual 'representation' (though, I have suggested, it is now misleading to call it that, see Section 5.3.7.1) is correct, then there is no sense in which experience can represent something as having a property when the subject has no idea what it is for something to have that property. The suggestion is a contradiction in terms. The fact that Shoemaker's account requires exactly this is closely related to the problem which Shoemaker admits he has, whereby he needs to rule out a very standard form of explanation of the representation relation, in order for his account to work (see the discussion of Section 4.6).

In the above account of colour qualia, we avoid any such problem. In basic – non-theoretically informed – experience, we are just aware of public colours. Even in such basic experience, we *have* colour qualia, but *having* them does not involve *being aware* that we have them. Instead it involves whatever is needed to fill out the 'space of reasons' story for the subject, so that the theorist can go from just being able to say *that* the subject is responding to red, say, to being able to say *how*. It is only in theoretically informed experience that the subject becomes *aware that* red is affecting them this way (some way, whichever way it is), and further, that red has the *property of* affecting them that way.

As such, unlike Shoemaker's account, the present account of qualia is compatible with (indeed, is premised upon) the claim that experience can only represent what the subject understands there to be; i.e. that what there seems to be, in the having of an experience, is fully capturable in the categories of the subject's understanding[108].

[108] Once again, the brief claim here represents an endorsement of the conceptualist viewpoint (e.g. McDowell, 1994), as against nonconceptualism (e.g. Evans, 1982; Peacocke, 2001). However, space and time preclude further discussion, in this thesis, of the very subtle issues involved in this debate, though some related points are made in the Appendix.

5.4.3 Pain Qualia

Could we extend such an account to pain? In giving the above account of colour qualia, I've typed qualia by what causes them. In other words, I've said that we are *affected redly*[109]. But red has two properties which let this account work the way it does. Firstly, it exists unperceived. Secondly, it is *not* such as to necessarily cause the effect it has, q_r; for perhaps the effects red has on me (in terms of my associations between red and other things, and in terms of affect) are the effects which green has on you (if so, we would have a case of *behaviourally detectable* inverted spectrum).

I think we would go very wrong if we try to find qualitative *pains* with either of these two properties, but all the same I think we can propose an account in which to be in pain (to have an experience with the quality of pain) is to be *affected painfully*, by *something*.

In the case of colour, I've suggested that the quale of red is the (behaviourally detectable) effect which red has on me – whichever effect that is. This seems entirely wrong for pains. We class things as pains (at least in what Lewis calls the *a priori* sense – Section 2.2.4; in what I would suggest is the *only* sense), in terms of the effect they have. Something is not a pain experience, in this *a priori* sense, if a subject responds with every sign of pleasure and satisfaction; and it is a pain experience, if the subject responds by trying to stop it, or mitigate it, in whatever way possible.

In trying to identify pain (pleasure, thirst, etc.) with such behavioural profiles, I might seem to be in danger of identifying pain with something much too simple – with patterns of behaviour which many *extremely* simple artefacts can show (Braitenberg, 1984). But note that Shoemaker (c.f. Section 3.3.6) found there to be some connection between our having conscious, experienced pain, and the ability we have to bring our *rationality* to bear, in avoiding pain. The Churchlands are less explicit about any specific connection between pain and rationality (see Churchland and Churchland, 1982 p.126) but they do require that a central part of the functional role of qualitative states in general (the example they use being the feeling of warmth) inheres in their link to conceptually structured states:

> "such as the belief that I have a sensation-of-warmth. If these sorts of causal relations are not a part of a given state's functional identity, then it fails to be a sensation-of-warmth on purely functional grounds. " (Churchland and Churchland, 1982 p.128).

[109] For more on the relation between the present view and classical adverbialism, see Section 5.5.

This link between pain and rationality is, I will argue, crucial. It is essential to ask how we can account for (or analyse) the *reason* which pain gives us, to act. According to any of the strong phenomenal realist accounts canvassed in Chapter 2, the reason why we call some particular (intrinsic) feeling pain, is because we are (contingently: it need not have been thus, on such accounts) motivated to respond in certain functional ways, when in states with this particular 'intrinsic quality'. This involves an extra level of indirectness which I believe is not needed, in the correct account of pain: we should identify the quality of pain with *the motivation to respond thus*, rather than with anything which is responded to, thus. But the 'motivation' in question is not just about involuntary responses; rather, it is the kind of *bona fide*, 'space of reasons'-level motivation which enables and *requires* us to bring our rationality to bear, in avoiding pain.

Thus, my proposal regarding the qualitative feel of pain is the following:

> The quality of pain is the affective modification of a space of reasons, which is such that the subject is motivated to respond aversively, (at least as if) to damage, or incipient damage, to a body part.

Therefore, this account is not as the Churchlands' account would be: one doesn't seek to jump out of the frying pan because one's experience has a certain quality, which one is motivated to respond aversively to. Nor is it as Shoemaker's account is: one does not seek to jump out of the frying pan because one has an experience which represents things as 'hurting', and where one is motivated to respond thus, to experiences which represent things thus. Instead, one's pain simply is the personal-level motivation to respond (at least as if) to damage, or incipient damage, to one's body parts. As such, the quale of pain, the feeling, is *not* the subject's reason for action on the account offered here; rather, the damaged body part (or at least seemingly damaged, at least seeming body part, in the case of illusion etc.) is the subject's most immediate reason for action, in such a state.

Again, as I emphasized for colour qualia, there has to be action in a space of reasons for there to be a mind at all. Then, for there to be pain, that pattern of action in a space of reasons has to be modified thus (i.e. in this particular way, which we *call* pain[110])

[110] I am not saying that our motivational structure itself has to fall into patterns which are all and only either pains or not pains. I am saying that splitting motivational structures into states which are pains and states which aren't is a rational way to respond to the motivational structures which we find, around here.

such that the creature simply is motivated to make those actions which *would* be rational, *were* the creature to be explicitly (i.e. thinking of it *as such*) seeking to mitigate damage to a body part. The conditionals in the preceding are, however, very important – the creature in pain *does not* need to have any concept of *damage* to a *body part*, as such, nor anything similar. Instead, its motivational structure simply needs to be changed such that *these* actions (the ones which are, *in fact*, the kind of actions which it would be rational to do, if explicitly trying to mitigate damage to a body part) are what the creature finds itself with reason to do, when in that state.

5.4.4 Are There Still Pains?

It has been widely supposed that, on any broadly adverbialist account[111], one simply has to deny the existence of pains, *qua* objects of perception. Certainly, traditional adverbialism was fighting against the view wherein pains are private, intrinsic, essentially mental 'objects' of awareness, whose nature is to be responded to painfully, and I have no wish to reinstate *those* pains.

But all the same, it seems to me to reduce the plausibility of any analysis of pain, if it has to say that there literally aren't *any* pains, *qua* objects of perception, in *any* sense. And equally, it seems to me to be quite possible to locate a place for pains, *qua* objects of perception, in the present account.

My proposal is that on this account, we should say that *the* pain – the thing in view for the subject, as a reason for action – is the (at least intentional) body part, which is sensed painfully. To clarify, I should point out that there is also a different meaning of the word *pain*, whereby pain (rather then *the* or *a* pain), is the painful state of the whole subject. I am not suggesting that normal language is entirely clean here. Just that there are at least these two senses of the word *pain*. At that on one of them – the former – *pains* are indeed *bona fide* objects of perception: they are body parts sensed painfully. And this enables them (pains; body parts sensed painfully) to be reasons in view for the subject (just as food is a reason for me to act, when I am hungry). Again, this is not to reinstate private sense data. The things I am talking about are normal, public, body parts, sensed painfully. I am, though, claiming that there is good reason to call these normal, public, things pains, *when they are sensed in this way*.

It is useful; it works; there are states which are pains, and there are states which aren't, and there are, of course, grey areas.

[111] Which this account is, very broadly, see Section 5.5.

5.4.5 Where Are Pains?

We also want to be able to ask (and answer) the question: where is my pain? Is it in my foot? In my tooth? Given the above analysis, then yes, pain is indeed in the relevant body part (in the relevant veridical case) in a perfectly valid sense: the pain (in the above sense of 'pain') is the veridically sensed part of my foot, with veridically sensed damage or incipient damage (although the subject need not sense the damage as 'damage', but must sense it as '[something to be prevented]'). Is the pain in *my mind*? Only in the same sense in which a perceived tree is in my mind: the pain (the body part, sensed painfully) is active as a reason, in my space of reasons. Equally, of course, pain (*qua* 'being affected painfully'), is now an introspectible property *of* my mind. Is the pain in *my brain*? No. Or rather, hopefully not: not unless I have felt, unpleasant damage to my brain, or it seems to me as if I do[112]. None of which, of course, is to deny that a lot of interesting low-level explanation about how I behave, when in pain, may refer to the detailed subpersonal states of my brain. On the other hand, a lot about how I behave, when in the fully veridical case of pain[113] is made true by the state of my foot (say), and by the state of the nociceptors in it[114].

5.4.6 Can Pains Exist Unperceived?

So, *these* pains are normal, public things (body parts), when sensed a certain way. Can they exist unperceived? Yes and no. Body parts can exist unperceived; that's the 'yes'.

---

[112] Apparently brains don't have pain receptors – i.e. damage to the brain cannot actually be felt in this way. Certainly, though, headaches can be intentionally *as if* 'that which is to be mitigated and prevented' is inside one's head. Perhaps that's the closest *we* ever get to a pain in our brain; in which case, we don't ever get all the way there. However, there does not seem, to me, to be any convincing, 'in principle' reason why some agent couldn't be thus; why brains *couldn't* have pain receptors, although there are occasional arguments in the literature attempting to explain why things couldn't be thus.

[113] That is, when sensing damage to a body part (in this painful way) when and because the damage is there.

[114] Including the actual, neurological c-fibres, which are in my extremities and not in my brain, contrary to common philosophical misconception (see Puccetti, 1977). Presumably, this misconception arose due to the fact that identification of pains with c-fibre firings was first made at a time when it was philosophically more popular than it is now to locate pains where they seem to be (and where c-fibres are), in the body parts sensed painfully; again presumably, the identification remained, as the philosophical trend moved towards locating pains (and therefore, mistakenly, c-fibres) in brains. Unfortunately, this is currently no more than a just-so-story, which could (and should) be confirmed or disconfirmed with an appropriate review of the relevant historical literature.

But body parts sensed painfully (i.e. pains as such) cannot exist unperceived; that's the 'no'. Nothing here brings back pains of the kind to be avoided: private mental objects which cannot exist unperceived. For clarity – to make it crystal clear that I am not trying to reintroduce that idea – perhaps I should go down the traditional adverbialist route, and say that there simply aren't pains. And, indeed, there simply aren't, *in the sense in which the adverbialist meant it*: there simply aren't *those* pains (those private, intrinsically awful, mental objects).

But the pains I am allowing are not those pains. And there don't seem to be any costs of allowing them, except for the possibility of misunderstanding. They are, I believe, a positive feature of the account; they are the reason for action, when in pain, from the subject's point of view.

I should note that I do not believe that I need to worry unduly if there are aspects of the English word 'pain' under which some bodily damage is sometimes correctly described as pain even when it is unsensed, or sensed but not sensed painfully (and this empirically can occur, under the influence of strong opiates for instance) (for references, see Aydede, 2005/2008 Section 5.1). For there is certainly only unsensed pain in such a case to the extent that the body part in question *would* be felt painfully, if only certain counterfactuals obtained.

As such, I should make clear that I am not here trying to accurately capture the exact sense of the English language word 'pain'; although I am, certainly, trying to capture accurately and consistently certain specific, central *aspects of* the meaning of that word.

5.4.7 The Different Feels of Pain

I have tried to avoid, in the above, talking about *the* quale of pain. Clearly, we can talk about the quale of a particular pain. There is, though, an outstanding issue to be addressed here, for my account. On accounts in which the feel of pain is determined by the role filler rather than the role, then the very same set of responses – pain – might feel one way in me, and another way in a silicon-based agent, say. That is to say, it is non-problematic for pain (the *a priori*, aversive state) to have more than one qualitative feel, on such accounts. The problem for my account is that it might seem hard or impossible for me to account for different feels of pain, at all. However, pain *clearly does* feel more than one way, even within a single subject. As the Churchlands' say:

> "Consider the wide variety of qualia wilfully lumped together in common practice under the heading of pain. Compare the qualitative character of a severe electric shock with that of a

> sharp blow to the kneecap; compare the character of hands dully aching from making too many snowballs with the piercing sensation of a jet engine heard at very close range; compare the character of a frontal headache with the sensation of a scalding pot grasped firmly. … [W]hat unites sensations of such diverse characters is the similarity in their functional roles … [including causing] involuntary withdrawal … [and] immediate dislike, … [being] indicators of physical trauma … . Plainly, these collected causal features are what unite the class of painful sensations, not some uniform quale, invariant across cases." (Churchland and Churchland, 1982 pp.125-126).

If the Churchlands are right, here, then my project is doomed: we cannot identify the feel of pain with the motivational structure when in pain, because (it would seem) there is only one *a priori* functional notion of pain here, involving only one *a priori* motivational structure, and yet there are many feels which are pains, in each of us.

I believe my account can cope with this. Commonly used examples of pain qualia span the whole range from sharp pains, searing pains, and dull throbbing pains through to itches and tickles. Is it really true that the functional role of all of these is the same? In fact, it seems quite clear that this is not the case, at least when we move as far from standard pains as to get to itches and tickles. These states by definition do indeed have their own *a priori* functional role: a Martian only has an itch, in the *a priori* sense, if that Martian has the urge to scratch it.

But is it really possible to make the same kind of move for sharp pains vs. searing pains, and vs. dull throbbing pains, and so on? I believe it is. For I think we would do well to look more closely than philosophers often do, at *how* these pains are classified in the first place. A *sharp* pain is the aversive reaction you get (as) to something sharp entering your skin. A *dull* pain is the opposite of sharp, the damage feels (at least *inter alia*) less precisely located than with a sharp pain. A *searing* pain is the aversive response you have (as) to your skin being seared (damaged by heat over an extended area). A *throbbing* pain throbs, the feeling comes and goes (or anyway modulates).

Thus, I do not think that the feel of a sharp pain, in me, could be the feel of a searing pain, in you. At least not unless one of us was motivated to remove the sharp thing (which sharp thing?) when the surface of our skin was being seared, and the other was motivated to mitigate or prevent the extended surface damage (which extended surface damage?) when something sharp entered our skin. The most appropriate response of a creature to something sharp entering it's skin is *not* the same as the most appropriate

response to more diffuse heat damage, say. And so on, for other types of damage and for other feels of pain.

So, whilst different pains *are* all (at least intentional, or 'as if') effects on the body which are to be responded to aversively, it seems to me quite possible that the different feels of different types of pain can indeed be captured within a) the differences in what, more precisely, seems to be the case, as regards the nature of the damage, when one is in one pain state rather than another, and b) the differences in what one is motivated to do about it, in the various cases.

5.5 Connections to Adverbialism and Direct Realism

Some readers will have seen my use of formulations such as 'affected redly' and 'affected painfully' (both used of thinking subjects) and 'sensed painfully' (used of body parts, sensed by thinking subjects), and will have worried that these are more than just reminiscent of traditional adverbialism. The two related worries would be: i) this account might simply *be* adverbialism, and/or ii) this account might be subject to the same objections which were responsible for the near terminal decline of adverbialism.

Other readers may be worried to see that I have indicated, at points, that I see this thesis as an endorsement of direct realism. For direct realism is often supposed to be (wilfully) anti-scientific.

Time and space preclude a fully detailed discussion of the many issues here, but I think I can say enough to explain briefly why I believe these various worries are misplaced.

Firstly, it is of interest to note that there seems to be some confusion in the literature as to whether traditional adverbialism was or was not a direct realist account.

For instance, Aydede says:

> "Direct realists … typically insist that such cases [as hallucination] should not be analyzed in terms of a perceiver standing in a certain perceptual relation to a private mental object or quality. Rather the analysis involves only one particular, the perceiver herself, and her being in certain sorts of (perceptual, experiential) states or conditions that are typically brought about under certain circumstances in which one genuinely perceives something. …
>
> This sort of analysis of experiences is sometimes known as adverbialism in the literature because in perceiving a red object one is said to be in a state of perceiving something "red-ly."" (Aydede, 2005/2008 Section 3.5)

On the other hand, Crane says:

> "The common kind assumption says that perceptions and hallucinations are states of the same fundamental kind, and hence it follows that this kind of state cannot be a relation to mind-independent objects [i.e. it follows that direct realism is false]. This inference is accepted, in one way or another, by the sense-data, adverbial and intentionalist theories." (Crane, 2005/2008 Section 3.4.1).

Having reviewed several early and more recent adverbialist positions (Ducasse, 1942; Chisholm, 1957; Sellars, 1975; Tye, 1984), I believe that the correct analysis here is that, whilst the core claims of adverbialism are indeed compatible with direct realism, adverbialism as it lived and breathed never was a direct realist thesis. Historically, adverbialists did not see themselves as questioning the common kind assumption, which is deeply entrenched in the sense-data theory which they rejected. Rather, they saw themselves as attempting to tame the common kind: to give a non-problematic analysis of it. However, the account of qualia given here (and the account of introspection[115], and of the mental level in general), does indeed reject the common kind assumption: there is no common factor between perception and illusion (or hallucination[116]) which is common to both, *and which is explanatorily more fundamental than either* (this is one of the most central points made by the originator of modern disjunctivism: Hinton, 1973).

I have already said something similar in Section 2.3.3, but I would like to briefly make explicit, here, that I do not think that anything in this latter claim requires us to avoid 'success-neutral' variants of words such as 'see' or 'experience' (as some direct realists have claimed). There is a perfectly valid sense of 'experience' in which I experience red *both* when I see red things *and* when I only seem to. Of course, on a direct-realist account, this success-neutral sense *means* no more nor less than: I successfully see red, *or* it is relevantly (in my behaviour, and for me) as if do. It is often complained that direct realists can say *nothing* about the success-negative cases (the 'bad disjuncts'), other then that they are subjectively like the success-positive cases (the 'good disjuncts') (c.f. Martin, 2006). But on the account offered here, as I will clarify

[115] In interpreting their account of introspection in this way, I would seem that I may differ from both Shoemaker (Chapter 4) and Sellars (Section 3.4.2.4).

[116] There is considerable disagreement as to whether illusion or hallucination is the correct case to contrast with perception, in expressing the central disjunctive commitment of direct realism (Byrne and Logue, 2009). I will not discuss these issues here.

shortly below, it really is possible to say *something* about the bad disjuncts, without introducing the fatal 'common kind'.

As well as there being a 'success-neutral' sense for 'experience', such a sense is also available (in ordinary English) for the word 'see' itself (as brief consideration of the situations in which 'seeing red' can be used indicates); in everyday usage, we slip between success-neutral and success-only variants as context requires, without even noticing.

What is being claimed here, however, is that there is nothing *explanatorily more fundamental* than either seeing (in the success-only sense) or illusion/hallucination (which are success-negative, partially and wholly, respectively) lurking within what is captured by the various success-neutral senses.

All of this cashes itself out in the (effectively, direct realist) claim of Sections 5.3.5 and 5.3.6 (and see also 5.6 below) that we see *objects*, in the first instance[117], and that we only come to know (be aware *of*) our qualia afterwards, as a more sophisticated act.

It should be noted that a frequently raised objection to adverbialism was its alleged inability to correctly analyse the sensing of a red triangle *and* a green square at the same time (the "many-property" objection, see, e.g. Jackson, 1977 p.59). In analysing such a case, the adverbialist can claim that an agent is sensing *redly* and *greenly* and *squarely* and *triangularly*, but it was never clear that adverbialism could account for the correct pairings. Jackson (1977 Ch.3) pushed this line of objection and other related points very forcefully. Authors such as Tye (1984) responded on behalf of adverbialism; but if anything, the artificiality of the moves required to try to repair the account looked to count against adverbialism. The problem for Tye seems to have come down to the problem of trying express all the right perceivings *within experience*, with experience conceived of as something definable separately from the world it is of.

But the present account needs no such moves. When one is sensing red-triangularly, at a certain place, this can be equivalently re-expressed by saying that one is behaving, or at least counterfactually would behave, in a certain way[118], towards a certain public

---

[117] Though sometimes (i.e. in the case of illusion and hallucination), we find our minds running – and ourselves acting – as if we see objects, even when we do not.

[118] Which way? 'Redly' and 'triangularly'. There are some subjective parts to each of these. For the analysis of the subjective parts of redly, see above. For triangles, these aspects include: do I *like* them? Do I like 'pointy' objects, or do I prefer smooth, rounded ones? Equally, there are some objective parts to each. This is a relatively trivial observation, for a triangle (mastery of what it is for something to be a

object. Now, if one is having an illusion or hallucination as of a red triangle, this is because *the behavioural profile is relevantly the same*, but either the object[119] is not there, or some aspect of one's relation to the object is not as it would be in the veridical case. As such, one doesn't need to create notation for sensing greenly at a certain point in the 'visual field' as Tye (1984, e.g. p.222) found himself forced to do, because one can instead just talk about acting (at least as if) towards a green object *there* (i.e. in the world).

It should be noted that the present analysis is also considerably more specific than traditional adverbialism. For traditional adverbialism never tried to spell out what it was to 'sense redly'. I am not here making an objection which adverbialism itself would have wished to reject: for the aim of adverbialism was to tame sense data by showing that they *could* be replaced without loss by a formulation in terms of modifications of the subject. As such, and as Jackson (1977 p.68) observes, adverbialism was always no more than a placeholder for a more complete theory of sensory feels, even in its own terms: whilst adverbialism showed (or aimed to show) the right *form* of an eventual theory, it didn't actually give a theory of that form.

To the best of my knowledge, direct realism has never been any more explicit than adverbialism as regards what feels are, and has often been less so (with the temptation being to deny that there are any such things to be known; to say that there is just the world of public reds and blues and stars and chairs, and that subjective qualitative feels are either no part of it, or can only be individuated in terms of non-relational public properties[120]).

I should also point out that it remains unclear to me whether (and if so, in what sense) direct realists have in fact endorsed the 'sensing redly' formulation (as Aydede claims, in the quote given earlier in this section). The two papers listed by Aydede as examples of "early direct realists" are papers by Ducasse (1942) and Sellars (1975). The latter paper is clearly (both from its title and content) an endorsement of adverbialism, and indeed each of these authors is more commonly listed as an adverbialist (e.g. Siegel,

---

triangle has non-optional behavioural aspects); but equally, you're not behaving as towards *red* at all, unless you do certain things. See Peacocke (1992) for one analysis of what things these are, and c.f. the Appendix herein, for my own analysis of how these points in Peacocke should be best understood.

[119] Which object? 'The' intentional one which the subject is *acting as if towards*.

[120] Pitcher (1970) is quoted by Aydede (2005/2008) as an example of someone making such a move.

2005/2008; Crane, 2005/2008; Lycan, 2000/2008). It is not clear that Aydede's quote does much to establish that direct realists, in the standard modern sense of the term, have endorsed this formulation.

More importantly, I am certainly not aware of any attempt within direct realism to say what 'sensing redly' actually comes to, any more than within historical adverbialism. Indeed it might look as if, were a direct realist to endorse this formulation, and to say what sensing redly comes to, that they would thereby be allowing the common kind back in, by giving an analysis of it.

On the present account, however, we can say what sensing redly (in the success-neutral sense) comes to, within a direct realist framework, without any such problems. Sensing redly is *behaving*[121] in a certain way: the way in which one does, in fact, behave when one encounters red objects (which has both subjective and objective aspects; c.f. footnote 118). What has this *behaviour* got to do with the phenomenal feel of red? I have argued that such at least counterfactual behavioural facts are both introspectible (Section 5.3.2) and include elements which are 'subjective' in the relevant sense (Section 2.2.7, and Section 5.3.1 above)[122].

This analysis of 'sensing redly' can, in a way, be seen as a 'common factor' between seeing and hallucination. But it is not the problematic common factor which the direct realist has to deny, for it is not that type of more fundamental common state which could be used to help explain (in the sense of Section 2.2.3) the nature of those states which it is in common between: one cannot use 'behaving as one does when one sees red objects' to help explain what 'seeing red objects' consists in.

Crucially, though, this is not to say that the project of analysis engaged in here is worthless: such analysis, if successful, helps us to get clear as to what we actually *mean*

---

[121] At least counterfactually behaving.

[122] To address one further worry: direct realism should *never* have been taken to be incompatible with the claim that there are subpersonal causal chains linking the subject to the world. Nor is direct realism incompatible with causal accounts at the personal level, such as Noë's account which I endorse in the Appendix. Of course, Snowdon (1980-81) is widely quoted as having shown that direct realism is incompatible with a causal account of mind (which, I would agree, would certainly count against direct realism). In fact, if read very carefully, it can be seen that Snowdon's paper does not touch causal accounts of the type Noë offers (nor does it claim to) (Child, 1992 has already made this point).

by mental terms[123], before (or, at least, in some sense, logically separably from) the process of explaining how such non-reductively characterised properties come to be instantiated in the physical world.

5.6 A Note on Order of Explanation

On a rather standard view, our awareness of the world is to be explained in terms of our more immediate acquaintance with broadly 'representational' states such as qualia. This is a classic 'Cartesian' view. In any relatively sophisticated version of such a view, it is not supposed that we are aware of (acquainted with) qualia *in the same sense* in which we are aware of (acquainted with) the world. This would be too evidently question begging. Rather, in a classic Cartesian account, we are supposed to be *more* directly acquainted with our qualia than we are with the world. It is supposed that we can be logically more certain of these 'inner' features of our mental lives than we can of any external features with which we are presented. Thus (once again, on the well known, classic, form of the view), it is supposed that an evil demon (Descartes, 1641) might be causing these internal states in us, such that we are systematically deluded about everything which is (apparently) public. All the same, the line of thought goes, we could not be deluded about the inner states themselves.

Nothing like this classic view survives here. I have already suggested (Section 4.3.1, Sections 5.3.5-5.3.6) that no real sense can be made of the suggestion that an agent is aware of state of affairs *x*, except to the extent that the agent can be shown to be rationally responsive to *x*[124]. As such, the most basic case of awareness is awareness of the world. Awareness of qualia comes afterwards. Qualia are introspected in the same way in which beliefs and desires are introspected: we can come to know them (be aware *of* them) only as and when we come to know ourselves as thinkers.

Nevertheless, we are indeed 'acquainted' with our qualia (exactly as with our beliefs and desires), in a certain sense (c.f. Chapter 2, footnote 41); for they are states of us, *qua* rational subjects as such. Therefore, they are exactly the right kind of states to be known

---

[123] By giving us a rich account of the inter-relation between such terms; McDowell (1994) can and should be read, at least *inter alia*, as being engaged in such a project (and he analyses – very largely successfully, in my own opinion – many other aspects of the mental on which I have touched barely, or not at all, herein).

[124] The agent must also be responsive to *x* under that description (whichever description it is); and responsive to *x* at least within an identifiable fragment of the space of reasons (c.f. Hurley, 2003).

by noninferential, single-step rational introspection (Chapter 3). This is indeed *something like* the supposed acquaintance with Cartesian qualia: for qualia, even on the present view, (and any other states of ourselves, *qua* rational minds as such) are inevitably 'ready to hand', ready to be known in this fundamentally first-person way.

All the same, there is something like a reversal of the classical order of explanation, here. In the classical (and still, I believe, prevailing) view, our acquaintance with our innermost properties is supposed to be primary. It is supposed, as it were, that we already know how to explain our first-personal acquaintance with the world, *in terms of* our acquaintance with our innermost properties. The difficult problem – perhaps even the hard problem (Chalmers, 1996) – is that of explaining the nature of, and nature of our acquaintance with, our innermost properties.

The account presented here argues that this gets things exactly backwards. This chapter has accounted for our phenomenal properties (and our knowledge of them), *in terms of* our acquaintance with the world. Access to (i.e., the ability to think about) qualia comes from two things: first, an understanding of the world (which any agent must have); second, at least some practical understanding of the subjective effect the world has on us, as rational agents (which is an understanding which only a more sophisticated agent can have). Crucially, though, if the arguments given herein are correct, this latter kind of understanding is also understanding of something which is as much a part of the publicly observable world as are the more obviously objective, public properties which we perceive. That is to say, our qualia *are* public, even if they are not *as* easy to tease out – not as manifest on the surface of things – as, say, the *public* colour red.

This can leave a sense of vertigo. For how are we to explain our acquaintance with the world, if not in terms of a more direct kind of acquaintance with inner states? The first step is to *characterise* the nature of our acquaintance with the world (c.f. footnote 123). If the arguments herein are correct, our acquaintance with the world (which is to say, our possession of a mind) must be characterised in terms of our action (and at least counterfactual action) within a[125] space of reasons. Having once *characterised* mind, we can then seek to *explain* how real agents come to have it.

Therefore, we come to see that our qualia (partially) constitute us, in the same way in which our beliefs and desires do: not at the subpersonal level, but at the personal level.

---

[125] …a fragment of 'the'… (c.f. Chapter 2, footnote 53)

To put it another way, belief, desire, affect, quale, etc. are the kind of terms which we must (or, at least, can) use, in *re-expressing* without loss what it is to be a person, whilst *remaining at the mental level*[126].

Our rational responsiveness *to the world* therefore becomes primary, in giving an account of our innermost mental life. I take this as a benefit, not a cost, of this view. For if things *are* so, mind suddenly starts to look much more naturalisable. I have tried to show how things can be so, consistent with our having a strong, and perfectly reliable, sense of there being something inner and subjective about our mental lives: inner and subjective in that these properties exist, are introspectible *as such*, and cannot be defined simply by saying *what* a creature sees, nor by adding *that* it sees it, nor even by adding a specification of the shared, public *words* it possesses, to describe what it sees.

5.7 Summary

The fact that there can be no truly private qualia follows, rather directly, from the analysis of introspection I have endorsed, on which our introspective conception of the mental is the very same conception as the public, third-person conception (Section 3.4.3), combined with the claim that qualia are, indeed, introspectible (at least in us, who seek to explain them) (Section 2.2).

If these claims are correct, then we must either deny that qualia exist (Dennett, 1988; Dennett, 1991), or we have to find some place for qualia within the public, behaviourally detectable mental level. I have tried to do this, by suggesting that qualia are 'the state of being affected that way', rather than 'that which affects us that way'. I have argued that the only thing *which* affects us 'that way' is the public property (green, say).

I have tried to flesh out many of the details of this account. In particular, by adapting certain aspects of Shoemaker's account of qualia, I have tried to make sense of the various different ways in which we can think, firstly, of public properties; secondly, of the phenomenal effects public properties have on us; and finally, of the (relational) property, which normal public properties have, of *having* such phenomenal effects on us.

---

[126] As such the relation between qualia, belief, desire, affect, perception, etc., and *mind* is seen to be parallel to the relation between freezing, boiling, viscosity, etc., and *water* (or 'wateriness': 'water' in that sense which remains neutral as regards whatever actually happens to instantiate this set of properties round here). C.f. Section 2.2.3.

In order to make plausible this identification of *being* affected in a certain way with qualia, I have argued that the way red affects me (in the relevant sense, *qua* property of a space of reasons, as such), is indeed introspectible, on the account of introspection which I have endorsed in Chapter 3. On the other hand, I have argued, no intrinsic property is thus introspectible. Indeed, Shoemaker now accepts this. Therefore, I would argue that the non-intrinsic properties which I have proposed have, on this basis, a better claim to be qualia than the intrinsic properties which have been so popular, historically, in analyses of qualia.

I have already argued that the proposed candidates for qualia can successfully naturalise some traditional intuitions concerning qualia. To further make it plausible that these properties are indeed qualia, I will argue in the next chapter that they do a good job (a much better job than Dennett, for instance, argues *can* be done) of accounting for various other intuitions which have lead people to suppose that there are intrinsic, or otherwise non-naturalisable, qualia.

6. *Reclaiming Qualia*

6.1 Introduction

The purpose of this chapter is to argue that the account of qualia just given can naturalise the intuitions lying behind claims that qualia are ineffable, intrinsic, private and infallibly or incorrigibly knowable. To the extent that this succeeds, it will bolster the argument that the features of a space of reasons which I have identified *are* qualia: those introspectible aspects of mental states which have traditionally been supposed to have these properties.

I phrase the above claims carefully (talking about 'naturalising intuitions') for I certainly agree with Dennett (1988; 1991 Ch.12) about *some* of his claims to the effect that nothing could have the above problematic properties, in *some* of the senses he discusses. It should be clarified, though, that my account of qualia is *not* Dennett's account. As will be seen, I don't go along with Dennett in all of his denials, and I certainly don't think that the whole theoretical framework surrounding qualia is so tangled that we must "get a new kite string" (Dennett, 1991 p.369). Instead, I think we *can*, with plausibility, identify features of our mental lives which would naturalise claims of ineffability, intrinsicness, etc.

I also disagree with Dennett in a perhaps more fundamental way. Dennett argues that his heterophenomenological account can show why "there *seem* to be qualia" (Dennett, 1991 p.372, emphasis added). But that is all; he thinks that there is no referent for these seeming properties; that they are a fiction, a part of a story we tell about ourselves. Qualia on my account aren't like that. They are a *bona fide* property of a subject's space of reasons – they are *not* fictions. They exist when not introspected. They can be known, in introspection. In a well-known turn of phrase which I quoted in the previous chapter, McDowell summarises the direct realist view of normal perception thus: "when one is not misled, one takes in how things are" (McDowell, 1994 p.9). So also with the introspection of qualia, on this account.

In order to discuss these various problematic properties, I will firstly look at Shoemaker's approach to defending a limited Cartesian thesis as regards self-knowledge of mental states. Though Shoemaker, rightly, shies away from saying that he has defended infallibility and incorrigibility as such, I think it is right to say that what he has

defended is sufficient to locate the source of our intuitions about infallibility and incorrigibility. It should be noted that Shoemaker himself has argued for this limited Cartesianism only as regards the more obviously intentional mental states such as belief and desire, and not qualia (indeed, it is unclear whether such an approach can be applied to qualia, on Shoemaker's own current analysis of them).

Next, I will address ineffability by means of a response to Dennett's most recent position paper on the knowledge argument. Dennett proposes that, if we wish to preserve a scientific account of phenomenal states, we must accept that full descriptive knowledge of such states is *sufficient* for an intelligent enough agent to come to know what it is like to be in such a state. I argue that Dennett is wrong about this, by his own functionalist, heterophenomenological lights. This argument has already been published elsewhere (Beaton, 2005). Here, I add the claim that Dennett has thereby mistakenly argued against a certain *bona fide* sense of ineffability, which can and should be preserved (and which, therefore, *follows from* a scientific account of qualitative states, rather than threatening any such account, as Dennett supposes).

Finally, I return to the remaining properties of a qualia which Dennett (1988) has tried to 'quine'[127]: intrinsicness and privacy. I argue that, whilst Dennett is right that we can find nothing intrinsic or private in any over-strong sense, we can naturalise various of the intuitions which have lead people to *say* that qualia are intrinsic and private.

6.2 Infallibility and Incorrigibility

6.2.1 Introductory Remarks

Infallibility and incorrigibility are two related issues in the area self-knowledge. Etymologically, one's knowledge of a mental state is infallible if it cannot fail (that is, what the subject says cannot be wrong) and is incorrigible if it cannot be corrected (that is, what the subject says is authoritative and final). The two notions are closely related. If there is a clear distinction between them[128], it is to do with direction of explanation.

---

[127] Dennett uses the humorous verb "to quine", from his own *Philosophical Lexicon*: "quine, v. To deny resolutely the existence or importance of something real or significant" (as quoted in Dennett, 1988). But, of course, as Dennett himself says, he is "not kidding" – he does indeed mean to deny that anything at all could have the properties which philosophers have traditionally ascribed to qualia (I would agree) and hence, that there are no qualia (for the reasons given here, I would disagree).

[128] Shoemaker makes no strong distinction between infallibility and incorrigibility, tending to use either infallibility alone, or phrases like "infallible or incorrigible" (Shoemaker, 1990 p.51; Shoemaker, 1988

When one thinks of infallibility, one thinks of a mental state (perception, sensation) and of a mechanism or process (in the most general sense) which infallibly leads to correct knowledge of that state. When one thinks of incorrigibility, the idea is that belief that one is in pain (say) entails that one *is* in pain: if one thinks one is, then one is.

The idea that our knowledge of our own mental states is infallible and incorrigible is widely perceived as "Cartesian": a working out of the Cartesian doctrine of transparency, that nothing can occur in a mind, of which that mind is not conscious[129].

Now as Evans puts it, whilst discussing related issues, a lot of philosophers have felt:

> "extremely suspicious of the idea of a judgement which is about something distinct from itself, yet which cannot be wrong" (Evans, 1982 p.229)[130]

And indeed, Shoemaker's aim is neither to defend such a view, nor to defend an account in which an agent *cannot be wrong*, about anything. All the same, Shoemaker does see himself as presenting a "limited … defense of Cartesianism" (Shoemaker, 1990 p.52). In particular, in (Shoemaker, 1988) and (Shoemaker, 1990), Shoemaker defends:

> "the Cartesian conception of the mind's epistemic access to itself – as a first approximation, the view that each of us has a logically "privileged access" to his or her own mental states, and that it is of the essence of mind that this should be so." (Shoemaker, 1990 p.50)

I think Shoemaker succeeds. I think he does show that it is of the nature of our mental states that we should know them, and know them correctly. I will explain (or at least, recapitulate, from Chapter 3) why this is. And I will argue that this is sufficient to *naturalise the intuitions* which lead philosophers to describe mental states as infallibly or incorrigible knowable. It is not that we cannot be wrong, nevertheless there *is* something special about self-knowledge; something different from our knowledge of non-mental facts, and of mental facts about others.

---

p.25) which indicate that he sees no important distinction. Indeed, in the index of his collected papers on self-knowledge (Shoemaker, 1996), "incorrigibility" is indexed "see infallibility".

[129] Cartesian transparency is quite different from the property which Moore ascribes to experience of its being "diaphanous" (Moore, 1903 p.450) (but also "transparent" at p.446). According to this very different, Moorean, version of transparency, mental states themselves cannot (or cannot easily) be grasped; rather, we 'see through' them to what they represent.

[130] Although I cannot fully endorse the nonconceptualist response to these worries which Evans puts forward on the same page of *Varieties of Reference*.

6.2.2 Self-Knowledge and Rationality

We have already examined Shoemaker's views on self-knowledge in detail, so we need only summarise them here. Shoemaker's essential claim is that, if one has mastered a concept which picks out a property of a space of reasons as such, then it is impossible to be both rational, and wrong in its self-application. I should remind the reader that Shoemaker in fact develops multiple separate lines of arguments for several separate, specific properties of a space of reasons as such (e.g. beliefs, desires); he himself never says that his arguments generalise to *any* property of a space of reasons as such, though I have argued that they do (Section 3.7).

Even though one cannot be *rational and wrong*, in self-ascription of one's mental states, one can be wrong. For every actual physical agent must be less than perfectly rational. But the upshot of all Shoemaker's arguments is that failure to noninferentially make the correct transitions (the ones that *would* legitimate claims of infallibility and incorrigibility if one *always* made them) *is* a failure of rationality.

Since all mental states are aspects of a space of reasons as such (or so I have argued) it follows that it is indeed *of their nature* to be known infallibly and incorrigibly. This is an ideal. But it is far from an irrelevant ideal. It is widely accepted that belief and desire are inherently rational states, even though creatures with beliefs and desires can be far from rational. My aim here – whether I succeed or not – is to make that same move equally as plausible for mind as a whole, including perception, sensation and phenomenal feel: these are *all* aspects of a space of reasons as such (just as are belief and desire), for all that real instances of mind fall far short of perfect rationality.

6.2.3 Self-Knowledge of Qualia

As I've already indicated, Shoemaker mainly discusses this "connection between special access to mental states and rationality" in the context of "intentional states like belief and desire", and he himself find it "less obvious how there can be such a connection in the case of our access to sensory states and, especially, to sensations such as pain" (Shoemaker, 1990 p.71). This is because (on Shoemaker's account, *but not on mine*), there remains an aspect of sensory states, to wit, their qualia, which *is* knowable (if indirectly) through introspection, but which is not an aspect of a space of reasons as such. I have argued that Shoemaker has to pay a very high cost for this, ruling out standard forms of scientific explanation, and possibly calling into question the very functionalist account which he claims to be endorsing (Section 4.6).

In the previous chapter, I have presented an alternative account which (perhaps surprisingly) retains recognisable descendents of some of the features of Shoemaker's most recent account of qualia but which – most importantly – identifies qualia themselves as *pure* properties of a space of reasons as such. As such, on this account, qualia (sensations, pains) quite naturally have exactly the same 'special access' properties as any other aspect of a space of reasons as such (including beliefs and desires).

Therefore, we can see what strikes me as a plausible motivation for claims that qualia (along with other aspects of a space of reasons) are infallibly or incorrigibly knowable by their possessor: to acknowledge that they are *not* infallibly or incorrigibly knowable (as we must) is to acknowledge that they (and we) fall short of the standards which, nevertheless, define them (and us, as agents).

6.3 Introduction to Ineffability

Something is ineffable if it is impossible to put it into words; the notion is usually associated with some sense of mystery, of the fundamentally inexplicable. Why are qualia supposed to be ineffable? Broadly, because we can know what it is like to have them, but, it is supposed, cannot capture what it is like in words. I believe this latter claim can be naturalised, in a way not threatening to physicalism.

I do not wish to naturalise the claim that qualia are fundamentally *inexplicable*. There is, though, a type of *inexpressibility* about the nature of qualia which I believe we can and must allow. The best way to introduce this type if inexpressibility will be to present recent work of Dennett's, in which he explicitly argues that allowing just this form of inexpressibility amounts to rejecting physicalism. In previously published work (Beaton, 2005) I have attempted to rebut Dennett's arguments on this issue. Here, I present the central portions of that paper[131]. The thrust of these arguments is that this type of inexpressibility, far from being a threat to physicalism, is entailed even by the strong functionalist form of physicalism which Dennett endorses. In the presentation here, I add the claim that this ineliminable inexpressibility can be plausibly seen as a naturalisation of a central aspect of the intuition that qualia are ineffable.

[131] As compared to the published version, I omit here: a review of previous responses to the knowledge argument; a brief analysis of phenomenal qualities, which is broadly compatible with, but superseded by, the account herein; and most of the concluding remarks, which are omitted in favour of the discussion of Section 6.5.

6.4 What RoboDennett Still Doesn't Know

6.4.1 Introduction

Mary, the colour-deprived neuroscientist, embodies perhaps the best known form of the knowledge argument against physicalism (Jackson, 1982; Jackson, 1986). She is a better-than-world-class[132] neuroscientist. Living in an entirely black-and-white environment, she has learnt all the physical facts[133] about human colour vision. She is supposed to be enough like us to be capable of having the sort of experiences that we would have on exposure to colour, but to be clever enough to know and understand the physical facts about her own colour vision, and to be able to work out all the relevant consequences of the facts which she knows.

The key premise of this form of the knowledge argument is that when Mary is finally released from her black and white captivity and shown coloured objects, she will learn something: namely, what it is actually like to see in colour. Indeed, in Frank Jackson's original paper, he takes it to be "just obvious" that Mary will "learn something about the world and our visual experience of it" (Jackson, 1982 p.130) on her release.

The following, then, is a simple version of Jackson's original knowledge argument, (all premises refer to Mary's pre-release epistemic status):

1) Mary knows all the physical facts about colour vision
2) Mary will learn something about what it is like to see in colour on her release

   Presumed corollary:

   Mary does not know all the facts about colour vision
3) Physicalism requires that if Mary knows all the physical facts then she knows all the facts

---

[132] Though perhaps not perfect, of which more later.

[133] I will use phrases such as 'physical facts', 'propositional facts', 'propositional knowledge' etc. more or less interchangeably to refer to the objective knowledge which Mary gains from black and white books, videos and so forth. Jackson states (or perhaps, claims) that after such an education a clever enough Mary could know "everything in *completed* physics, chemistry, and neurophysiology, and all there is to know about the causal and relational facts consequent upon all this" (Jackson, 1986 p.291). In this context, Alter has talked of the "discursively learnable" facts (Alter, 1998 p.50 and passim) and Churchland talks of those facts which are "adequately expressible in an English sentence" (Churchland, 1989 p.144). I am happy to accept the standard set-up of the knowledge argument, in which such knowledge exists, is learnable by a clever enough student via the route described, and is, further, contrastable with knowledge such as "red is like *this*" which Mary does not gain (at least not directly) from her black and white book learning.

Conclusion:

Physicalism is false

Premise 2) both implies and is implied by what I will call 'the Mary intuition'. This is the intuition that Mary, in the circumstances described, will still learn something on first seeing a coloured object (equivalently, that there is something that Mary, in the circumstances described, does not yet know, namely what it is like to see in colour). Paul Churchland has argued persuasively (Churchland, 1985; Churchland, 1989; Churchland, 1998) that every possible form of Jackson's argument requires some equivalent of premise 3) above which only appears to go through because of equivocation on two different senses of the word "knows". If he is right, the argument does not go through, and 'the Mary intuition' is *compatible* with physicalism.

This is one of two possible physicalist responses to the knowledge argument. The other major approach is to argue for the falsity of premise 2): to argue that the Mary intuition is incompatible with physicalism, and is false. Such a response amounts to a defence of the validity (though not the soundness) of the knowledge argument: it implies the claim that there is indeed some valid reasoning which shows that Mary's learning something new is incompatible with physicalism, exactly as Jackson originally claimed.

Jackson has now retracted his own knowledge argument (Jackson, 1998a; Jackson, 1998b; Jackson, 2003). It looks at first as if he has endorsed the second kind of response just mentioned. In his initial retraction, he stated that "*after* the strength of the case for physicalism has been properly absorbed" (Jackson, 1998a p.vii), one is "reluctantly" (Jackson, 1998a p.vii) led to conclude that "The redness of *our* reds can be deduced in principle from enough [information] about the physical nature of our world despite the manifest appearance to the contrary that the knowledge argument trades on" (Jackson, 1998b pp.76-77). More recently Jackson has stated that "physicalists are entitled to reject" (Jackson, 2003 p.9) "[t]he epistemic intuition that founds the knowledge argument [, …] that you cannot deduce from purely physical information about us and our world, all there is to know about the nature of our world because you cannot deduce how things look to us, especially in regard to colour." (Jackson, 2003 p.2).

That certainly sounds as if Jackson is rejecting his own premise 2, and saying that you *can* work out from enough information about us and our world, what it is like to see red. But this is not what Jackson is saying. He still accepts the truth of what I have called the Mary intuition; he still believes that Mary "would learn what it is like to see

red" (Jackson, 2003 p.3) on her release (indeed he continues to treat this as an obvious fact, in need of no defence). Instead, Jackson is rejecting the epistemic intuition (which he previously endorsed): he now accepts that 'learning' how things look is not a matter of "learning something about the nature of the world" (Jackson, 2003 p.3). On Jackson's revised account, Mary will learn no new fact about the world, but will instead gain a new kind of 'representation'; one with the right properties to account for the "immediacy, inextricability, and richness" of seeing red, and one which additionally grants her the ability to "recognise, imagine and remember" red (Jackson, 2003 p.26). As Jackson himself points out (Jackson, 2003 p.28), he has thus come to adopt the ability-based rejection of his knowledge argument originally employed by Nemirow (1980) and Lewis (1983). Jackson's revised position effectively leaves the knowledge argument exactly where Churchland left it, with true premises, but nevertheless invalid due to equivocation on two senses of "knows".

If we accept these arguments, can we consider interesting discussion on the knowledge argument closed? Apparently not, for the above, seemingly straightforward, physicalist consensus – now including Jackson himself – remains radically different from the position held by Daniel Dennett (who is, of course, another die-hard physicalist).

6.4.2 RoboDennett

Dennett's position is made clear in his new paper on the subject, "What RoboMary Knows" (Dennett, 2005b)[134]. For Dennett, "most people's unexamined assumptions imply dualism" (p.107; for which, in context, read "the Mary intuition is incompatible with physicalism"). The explicit objective of Dennett's new paper is to show that the Mary intuition is an anti-physicalist confusion. He aims to demonstrate – for the benefit of those philosophers who doubt that it can be done – how Mary "*figures out* exactly what it is like to see red (and green, and blue)" (p.122).

---

[134] The paper from which the present discussion is extracted was originally written in response to an online version of an article of Dennett's, which was due to appear in a collection of papers on phenomenal knowledge (Alter and Walter, 2006), and which is currently available online at http://ase.tufts.edu/cogstud/papers/RoboMaryfinal.htm. The current, and the published, versions of this discussion now take their quotes from that version of Dennett's paper which appeared (in slightly modified form) as Chapter 5 of "Sweet Dreams" (Dennett, 2005a). As such, all quotes from Dennett within Section 6.4 refer to Ch.5 of "Sweet Dreams" unless otherwise indicated.

But why should Dennett believe that most people's unexamined assumptions imply dualism? Or that philosophers need to understand why the Mary intuition is false in order to understand how physicalism can be true? He must still believe that there is some logically valid form of the knowledge argument, implying a fundamental incompatibility between the Mary intuition and physicalism, despite all the arguments to the contrary. Has Dennett simply missed the equivocation on "knows" from which, Churchland has claimed, all forms of premise 3) suffer? The actual situation is more complex than that, and more interesting.

The explicit aim of Dennett's new paper is to show that Mary will necessarily be able to come to know what it is like to see in colour, if she fully understands all the physical facts about colour vision. I believe we can establish that Dennett's line of reasoning is flawed, but the flaw is not as simple as an equivocation on "knows". Rather, it goes to the heart of functionalism and hinges on whether or not Dennett is correct to claim that there is "no fact of the matter" (Dennett, 1988; Dennett, 1991; Dennett, 1994, etc.) about what subjective experience consists in.

6.4.3 The Blue Banana Alternative

Dennett's previous major position statement on the knowledge argument occurred in his book "Consciousness Explained" (Dennett, 1991 pp.398-401). There, he first outlined in print what he believes to be a perfectly legitimate alternative ending to the Mary story. Instead of experiencing "surprise and delight" (Graham and Horgan, 2000 p.72) on being released from her room and first seeing coloured objects, something quite different happens. Mary's captors decide to trick her, and the first coloured object they allow her to see is a blue banana. Dennett doesn't explicitly state as much, but presumably Mary's captors are expecting Mary to say to herself something like, "Ah, so that is what yellow looks like!" However, what Dennett does say is that Mary isn't fooled for a moment, she takes one look at the blue banana and says, "Hey! You tried to trick me! Bananas are yellow, but this one is blue!" and further "I was not in the slightest surprised by my experience of blue (what surprised me was that you would try such a second-rate trick on me)" (Dennett, 1991 pp.399-400).

Dennett states that students and professional philosophers alike have had considerable problems with his alternative ending to the story (Dennett, 2005b p.106). But what exactly is this alternative ending supposed to indicate? Is he seriously trying to claim that Mary has "figured out" what it is like to see in colour without ever having seen

anything coloured? That is, of course, exactly what he is trying to claim. And he is not just stating that Mary will know enough about her own physical reactions to colour to be able to recognize them when they first occur, and so work out which colour she has seen. He is, rather, taking the following much stronger position: that knowing as much about your own reactions in advance of the fact as Mary does is logically equivalent to knowing what it is like to see colour in advance of the fact. He explicitly states that he knows of no "distinction … between knowing "*what one would say and how one would react"* and knowing "what it is like". If there is such a distinction, it has not yet been articulated and defended, by [anyone] … , so far as I know" (Dennett, 2005b footnote 3).

To many, of course (even to those who hold to the truth of some form of physicalism) this current, clear and explicit statement of position by Dennett will itself seem extreme. This is why he has felt compelled to return to the fray, and to attempt to "convince a few philosophers" (Dennett, 2006) that his position might be correct after all.

6.4.4 Introducing RoboMary

Dennett's chosen weapon for his final attack on the knowledge argument is RoboMary, a perfected robot neuroscientist. Dennett uses RoboMary because he needs to discuss the physical details of her behaviour and thought processes at a level of detail not currently available to human neuroscience. Using RoboMary he hopes to show, by analogy, how a human-like Mary could also come to know what it is like in advance of the experience.

I am happy with this approach, and agree with Dennett that a physicalist account of what is really going on in the Mary thought experiment will require a discussion of the physical details of the 'agent' under discussion. As Dennett says:

> "If materialism is true, it should be possible ('in principle!') to build a material thing – call it a robot brain – that does what a brain does, and hence instantiates the same theory of experience that we do." (Dennett, 2006)

and further:

> "Those who rule out my scenario as irrelevant from the outset are not arguing for the falsity of materialism; they are assuming it" (Dennett, 2005b p.125).

Dennett wants to make sure that RoboMary is a well constructed and well labelled "intuition pump". He succeeds admirably. In fact, once I have summarized here Dennett's key "knobs" and "settings" for RoboMary, she will make an ideal subject on

which to attempt some "cooperative reverse-engineering" (Dennett, 2005b p.122) of my own.

There are two major models of RoboMary, either of which, it is argued, can come to know what it is like to see in colour in advance of the experience. As Dennett outlines these two versions of RoboMary he considers and refutes many possible objections to his account. On many, indeed most, of these points I am fully in agreement with Dennett. Therefore I will only give an outline of the key facts about RoboMary, omitting the several objections to his story that Dennett successfully addresses.

6.4.5 Unlocked RoboMary

The basic RoboMary model is (for reasons presumably lost in the mists of sci-fi time) a standard Mark 19 robot. The easiest thing to do will be to quote directly the key points from Dennett's story about her:

> "1. RoboMary is a standard Mark 19 robot, except that she was brought on line without color vision; her video cameras are black and white, but everything else in her hardware is equipped for color vision, which is standard in the Mark 19.
>
> "2. While waiting for a pair of color cameras to replace her black-and-white cameras, RoboMary learns everything she can about the color vision of Mark 19s. She even brings colored objects into her prison cell along with normally color-sighted Mark 19s and compares their responses – internal and external – to hers.
>
> "3. She learns all about the million-shade color-coding system that is shared by all Mark 19s.
>
> "4. Using her vast knowledge, she writes some code that enables her to colorize the input from her black and white cameras (à la Ted Turner's cable network) according to voluminous data she gathers about what colors things in the world are, and how Mark 19s normally encode these. So now when she looks with her black-and-white cameras at a ripe banana, she "sees it as yellow" since her colorizing prosthesis has swiftly looked up the standard ripe-banana color-number-profile and digitally inserted it in each frame in all the right pixels.
>
> "5. She wonders if the ersatz coloring scheme she's installed in herself is high fidelity. So during her research and development phase, she checks the numbers in her registers (the registers that transiently store the information about the colors of the things in front of her cameras) with the numbers in the same registers of other Mark 19s looking at the same objects with their color camera eyes, and makes adjustments when necessary, gradually building up a good version of normal Mark 19 color vision.
>
> "6. The big day arrives. When she finally gets her color cameras installed, and disables her colorizing software, and opens her eyes, she notices . . . . nothing. In fact, she has to check to

> make sure she has the color cameras installed. She has learned nothing. She already knew exactly what it would be like for her to see colors just the way other Mark 19s do." (pp.122-125)

For what it is worth, I buy into this story. There don't seem to me to be any interesting reasons why RoboMary can't do what Dennett claims, above, that she can do. And if she can indeed do the above then she would indeed come to know what it is like to see in colour in advance of the experience. But an objection that Dennett considers concerning his step 4 is the crucial one, in terms of relating the story of unlocked RoboMary to the story of Mary. The question is, is unlocked RoboMary cheating or not when she writes directly to her colour coding registers? Perhaps, as Dennett himself says, RoboMary's colorizing system is simply the "robot version … of transcranial magnetic stimulation" (p.124): cheating in the sense of using a non-surprising way of coming to know what it is like, which doesn't truly involve deducing what it is like from the facts one knows. Or perhaps we should accept that "RoboMary is entitled to use her imagination, and that is just what she is doing – after all, no hardware additions are involved" (p.124).

Dennett is happy to vary this setting in both directions. For reasons related to the above point about imagination, my understanding is that Dennett thinks there is no truly principled reason to rule out even this unlocked version of RoboMary as a counter-example to the Mary intuition. (I will argue below that there is, in fact, a principled reason to rule that unlocked RoboMary's route to coming to know what it is like is cheating.) Nevertheless Dennett is happy to take on board this objection, and to consider next a much more challenging version of the RoboMary story.

6.4.6 Locked RoboMary

Following Dennett, "let's turn the knob and consider the way RoboMary must proceed if she is prohibited from tampering with her color-experience registers" (p.126). The use of a robot instead of a human in the thought experiment once again pays dividends. As Dennett says, we have no idea how "Mary could be crisply rendered incapable of using her knowledge to put her own brain into the relevant imaginative and experiential states" (p.126), but we can easily describe something equivalent for RoboMary. We can put a software system in place which automatically converts all the colour values in Mary's visual array to black and white (or rather, greyscale) values before any further processing takes place. Now let's put unbreakable software security on this system.

Suddenly RoboMary really can't "imagine" herself into any normal colour vision state. She can't even create colour 'phosphenes' (one objection to the original Mary story) by any robot equivalent of rubbing her eyes. The only way her colour registers can ever come to contain any usable colour values is for the software security system to be disabled which, let us assume, requires a hardware change and so can be treated as unambiguous cheating.

Surely then there is no way for RoboMary to deduce what it is like to see in colour, is there? Oh yes there is, says Dennett:

> "This doesn't faze her for a minute, however. Using a few terabytes of spare (undedicated) RAM, she builds a model of herself and *from the outside, just as she would if she were building a model of some other being's color vision*, she figures out just how she would react in every possible color situation." (p.126)

This is supposed to be pure heterophenomenology. For Dennett, there can be no distinction between the full facts about "what one would say and how one would react" and the full facts about "what it is like". Thus, if RoboMary can indeed build such a model, she can indeed come to know what it is like. QED.

But the preceding is a reconstructed abbreviation of Dennett's argument. Let's follow the actual details of the story which Dennett gives. Rather than mix and match direct and indirect quotation, I will paraphrase this section of Dennett's argument (pp.127-128). Imagine, says Dennett, a situation in which (locked) RoboMary is shown a ripe tomato. She can see it and touch it and find out all about its bulginess and softness. She can also consult an encyclopaedia to find out exactly what shade of red it would be, if only her colour registers were unlocked. RoboMary will react in various ways to this stimulus, resulting in some complex, internal, grey tomato experiencing state, state A. But at the same time, she can feed into her internal model of herself the true red colour values which she knows she would have seen if her colour vision equipment was normal for Mark 19s. So her model will go into a different complex state, a red-tomato-experiencing state, state B. This should be fine: the model RoboMary doesn't have to be 'locked', just because RoboMary is. She knows all about how she would work if she was not locked, and so she should be able to build and operate an unlocked model just as Dennett describes. So now, returning to direct quotation, locked RoboMary compares state A with state B and:

> "being such a clever, indefatigable and nearly omniscient being – makes all the necessary adjustments and *puts herself into state B.*" (p.128)

Dennett is at pains to point out that state B really isn't an illicit state in the sense in which direct tampering with colour registers is an illicit state. State B is the state that Mary would have gone into if she had had the colour experience, even though she hasn't in fact had it: she isn't making herself experience colour (cheating) she is making herself be as she would be if she had experienced colour (not cheating)[135].

I am prepared to buy into this story, too. I accept that locked RoboMary could find such a state and put herself into it. But I don't accept that RoboMary has told us about what *must* be true of an agent in the epistemic situation of pre-release Mary; I don't accept that she is not cheating.

6.4.7 What Physicalism Requires

For convenience, let's recap, with a quick and simple version of the knowledge argument:

1) Mary knows all the physical facts
2) Mary does not know what it is like
3) Physicalism says that if you know all the physical facts then you know everything

Conclusion:

Physicalism is false.

How should a physicalist respond?

Most physicalists, including Jackson (now), Nemirow, Lewis and Churchland have been prepared to accept that there is some distinction between the type of knowledge which Mary has, pre-release, and the type of knowledge which she gains on her release. Some physicalists have argued that Mary gains a new ability but does not thereby come to know *any* fact – not even an old fact in a new way; other physicalists have argued that Mary gains a new type of knowledge of an old fact. The important point here is that *both* these responses accept that it is possible for Mary to know all the physical facts and, at one and the same time, not to know what it is like.

Surprisingly, perhaps, even Dennett accepts this.

In either version of Dennett's story, RoboMary has to *do something* in order to come to know what it is like. She either has to adjust her colour registers, or she has to work

[135] Dennett draws an instructive analogy here with Swamp Mary (another character whom Dennett introduces, whilst suppressing his "gag reflex" and "giggle reflex"; p.120). I won't go into these details here, but I think that his point goes through.

out some special state, *state B*, and put herself into it. She's never just *automatically* in state B, as soon as she's finished learning all the facts. So pre-release RoboMary is like this: if you ask her what it is like to see ultramarine, say, she says "I don't know, but I can work it out. Hold on a minute [or a second, or a picosecond] … Ok, there we are! Now I know."

This is just what should be expected, on the account of qualia which I am attempting to defend, in this thesis. To know what it is like to see ultramarine is to *be* affected in a certain way by ultramarine, and not just to know what being affected in that way would consist in. This is why Dennett's line of argument does not admit of any simple mapping onto traditional responses to the knowledge argument, as discussed above. For, even on Dennett's own account, there is no claim that these two states (knowing what it is to be affected a certain way; being affected that way) are *the same*. Rather, he thinks that physicalism requires that Mary be able to *make the transition* from one state to the other, or as he might put it, be able to *work out* what it is like to see red from all her factual knowledge; that believing otherwise is an anti-physicalist confusion. Why? It appears that the version of the knowledge argument which Dennett must be using is the following:

1) Mary knows all the physical facts
2) Mary cannot work out what it is like
3) Physicalism requires that if you know all the physical facts, you can work out what it is like

Conclusion:

Physicalism is false.

If you wish to preserve physicalism under this argument, and you accept premises 1 and 3, then you have to reject premise 2. Conversely, if you accept premise 1, and you wish to preserve physicalism, you still have no reason whatsoever to reject premise 2 unless you think that premise 3 is true. That is, there seems to have arisen a physicalist consensus that 2 is compatible with 1; my aim here is to endorse this consensus, and then to draw some further conclusions from it. But for now, we are looking at Dennett's reasons for not joining this consensus. We have now got as far as establishing that he thinks he cannot, *because* he thinks premise 3 is true. So now we need to examine the logical status of premise 3 in the above argument in more detail, in order to see how Dennett's RoboMary ought to impact on our response to this form of the knowledge argument.

To be clear about the logical status of premise 3, we have to think about what *might* and what *must* be true of agents who know as much as Mary.

I am not particularly interested in what might be true of agents in worlds where mental facts float free of physical facts. Let's talk only about universes such as ours (I hope) in which all facts supervene on the *physical* state of the universe (the state as it would be described, if we had that much knowledge, in terms of the completed laws of physics)[136]. We can then ask, what might and what must be true of agents who know as much as Mary, in such purely physical universes? I will say that some predicate is *necessarily* true of such an agent if it must be true of every agent which could possibly be built, consistent with the laws of physics, who knows as much as Mary. I will say that some predicate is *possibly* true of such an agent, if that predicate *can* be true of an agent who knows that much – consistent with the laws of physics – but doesn't have to be.

Thus, I would claim, it is *necessarily* true that Mary can work out what 2 + 2 comes to, but it is only *possibly* true that (for instance) Mary's brain has built in to it a transcranial magnetic stimulation machine, which she can operate at will, which results in coloured visual phosphenes.

Now, for Dennett's arguments to work, it needs to be the case that Mary can *necessarily* work out what it is like to see red[137]. If she can only *possibly* work this out (if some agents who know that much can work it out, but some other agents who know that much cannot), then Dennett's argument is flawed. At best, RoboMary might lead one to accept that belief in the Mary intuition is belief that Mary has one physically possible type of architecture rather than another, which is not an anti-physicalist position at all. At worst (for Dennett's current position) there may be a good reason to believe if you extend human reasoning in the most natural way, to end up with an agent with Mary's abilities and knowledge, then you end up thinking about an agent with the 'can't-work-it-out' architecture. If this is so, the Mary intuition is better than equally as physical as its denial: it is the correct intuition to have had about Mary all along.

---

[136] This supervenience relationship means, simply, that you can't change any fact (of any type) without changing some physical fact.

[137] We are talking about an A-grade student here, who will not miss, or misunderstand, consequences of what she knows. As such, and as I will argue in detail below, Mary does – *necessarily* – have the ability to get very close.

6.4.8 RoboDennett

I have argued that the key question, which determines whether or not the Mary intuition is compatible with physicalism, is whether or not an agent who knows as much as Mary can necessarily use that knowledge in order to come to know what it is like to see red, if she so chooses.

Here, I will argue that there is nothing in the set up of the knowledge argument which *requires* that Mary be able to do what Dennett's RoboMary does. On the contrary, I will aim to describe a perfectly physically well defined robot agent who can know quite as much as Mary, or RoboMary, but who remains genuinely unable to come to know what it is like, despite mastering all the abilities that Mary is granted by the first two premises of the knowledge argument (that is to say, the knowledge Mary has, and at least the *potential* to come to 'know what it is like' in the way in which we do).

In order to regiment the discussion we need, finally, to be clear about what we mean by cheating in the context of the knowledge argument. I suggest that the correct way to proceed is as follows:

> When considering an agent trying to achieve what RoboMary achieves, in the context of the knowledge argument, the agent should be considered to be cheating if it uses abilities other than those entailed by the hypotheses of the knowledge argument.

I have already suggested, in the introduction to this paper, what these abilities are. The agent in question *must* be quite like us, for she must be capable of knowing what it is like to see red in the same way in which we do. Premise 2 requires this – we all grant that, after normal exposure to red, Mary will know what it is like to see red in the same way we all do.

On the analysis of qualia proposed earlier, this means that Mary must be able to come to act towards visually presented colours within her space of action for reasons. Putting the same point in more engineering oriented terms, it means that it must be possible to for the low-level colour responsive 'circuitry' in Mary to become appropriately recruited into her reason respecting behaviour. But what does 'appropriately' mean, here? It means, at least[138], that Mary becomes able to identify colours noninferentially, which is to say: not, at the whole subject level of description of her actions, in virtue of her recognising something else.

---

[138] For a little more on what is involved, see Section 6.4.10.

Premise 1, on the other hand, requires that Mary's abstract reasoning powers be much better than ours. She knows everything there is to know about how her own colour vision works. Moreover, she can work out any relevant consequences of what she knows. We should be wary of granting Mary perfect reasoning powers, but I don't believe that we need to. What we need to allow is that anyone trying to show just what Mary can do, can help themselves to any particular reasoning process, by Mary, based on her vast knowledge – but *only* in terms of reasoning from propositionally expressed knowledge to more propositionally expressed knowledge. This, of course, is the key move, but it does not, yet, establish the falsity of Dennett's position, for, as we will see, there are very good reasons (quite the best reasons, in fact) for thinking that these abilities alone *are* sufficient for creating a *bona fide* state of knowing what it is like.

Using the above limitations, I will define a new robot which I will name RoboDennett. RoboDennett is, of course, extremely intelligent, and he knows an awful lot – quite as much as Mary, or RoboMary, in fact. The only difference between RoboDennett and RoboMary (if indeed there is a difference) is that RoboDennett has no abilities which are not necessarily granted to him by the premises of the knowledge argument.

RoboDennett is, I suggest, the agent whom we should have been imagining all along, in the context of the knowledge argument. If the Mary intuition is true, of him, then the Mary intuition is not just compatible with physicalism, it is the *correct* intuition to have about someone who starts off like one of us, and who is only changed as little as possible in order to come to know as much as Mary knows. This remains so *even if* there are other physically possible agents (such as Dennett's RoboMary, for instance), who *can* use all their knowledge to come to know what it is like prior to exposure to colour.

But my argument does not depend crucially on whether RoboDennett is 'more' like us than RoboMary. My basic point is that if RoboDennett, with all and only the abilities an agent *must* have, in order to be an agent such as the one under discussion in the knowledge argument, cannot work out what it is like, then the knowledge argument does not threaten physicalism in the way in which Dennett takes it to. RoboDennett, of course, is very like RoboMary. RoboMary certainly *has* the abilities which I have granted to RoboDennett. The only substantive question is whether or not she exceeds them.

6.4.9 RoboDennett and Unlocked RoboMary

I said before that there were principled reasons for declaring that unlocked RoboMary was cheating. You will recall that she works our what colour values should be in her low level colour circuitry, and then simply puts them there. Of course she can work out what the colour values should be, but there is no reason to think that we humans have the ability to configure our low level colour processing circuitry in the way unlocked RoboMary does, just by thinking about it, in advance of any exposure to colour. More pointedly, I believe that there is no argument which says that an agent who knows as much as Mary somehow automatically gains the ability to do this. Apologies for having only shifted the burden of proof, but I think I have shifted it quite far. Lacking an argument for the *necessary* presence of this additional ability, unlocked RoboMary really was going beyond her legitimate powers of imagination, she was doing something which we cannot do with our imaginations, and something which increasing our reasoning powers up to the level of Mary's would not enable us to do. She was cheating.

6.4.10 RoboDennett and Locked RoboMary

As I've already indicated in Section 6.4.7, I don't think that Dennett has somehow entirely missed the central point I am making. He is less explicit about it than I have tried to be, but he recognises that what he actually needs to show is that any agent who has mastered all Mary's knowledge must necessarily be able to use that knowledge to come to know what it is like. Rather, I think this is precisely what he believes he has shown, using locked RoboMary. As we look in detail at Dennett's reasons for believing that the Mary intuition is fundamentally unphysical, we will see that what locked RoboMary does is indeed, by Dennett's lights, a completely general route to coming to know what it is like, a route which would be available to any agent who knows as much as Mary and can work out the consequences of what she knows.

For most of the steps on locked RoboMary's path to enlightenment, I am in full agreement with Dennett. Nevertheless, I believe that RoboMary does not correctly represent the entailments of physicalism. The final step (and only the final step) which locked RoboMary takes is a perfectly physical move, but it is a step which Dennett should not have allowed her, for it is a step which is not available to RoboDennett.

It is no accident, given Dennett's heterophenomenology, that locked RoboMary's route to coming to know what it is like involves working out exactly what she would

say and how she would react on exposure to colour. What she has actually done, just by thinking hard, is to create a simulation of herself. And, I will argue, this is a step which RoboDennett *can* take[139], even without the explicit provision of spare, undedicated RAM and processing power.

Imagine that you, yourself, knew everything about how a pocket calculator works (not the atoms or the quarks, just the registers, the CPU instruction set, and the relevant connections to the keys and the LCD display). Is it plausible that, once you knew all this, you could do without a pocket calculator? Of course not, for you are too human. You would make mistakes sometimes, as you tried to work out what the calculator would do, and even if you were very careful, and did get the answers right, you would be much slower than the calculator.

But to think that RoboDennett would still need a calculator, once he had put his mind to understanding one, is indeed to make precisely the mistake which Dennett accuses us all of making with regard to Mary. For RoboDennett is much better than us (as, indeed, are RoboMary, and Mary too). Once he has put his mind to understanding a pocket calculator, it would be obvious to him what the result would be of calculating sin(37π/5)^6 (for instance)[140]. That is to say, these agents are good. Very good. And, crucially, they are all supposed to be equally good even at the vastly more complex task of understanding themselves.

Are we still within the bounds of sense here? Is it possible to make any meaningful statements about an agent who is supposed to be a) in some relevant way, human-like, but b) to know as much, and be as good at using that knowledge, as Mary, RoboMary or RoboDennett are supposed to be? Yes, I believe so, though we have to steer carefully in these waters.

In the example of the calculator, above, RoboDennett's understanding of the calculator becomes good enough for him to *do away* with the actual calculator if two crucial conditions obtain:

---

[139] Which means: is a step which any agent under discussion in the knowledge argument *necessarily* can take, RoboDennett being the agent who can do all and only what such agents necessarily can do. (There are certainly complications here, perhaps such an agent necessarily can do *either* A and B, *or* C and D, but doesn't have to be able to do both; and one can think of similar possibilities of arbitrary complexity – I am not aware of any such complications actually being relevant to the present argument, however.)

[140] It's approximately 0.74, and I don't happen to know how many decimal places were on the calculator which RoboDennett was thinking about.

i. His understanding is so good that it is functionally isomorphic to the relevant level of organization of the calculator itself.

ii. He can operate this functionally isomorphic understanding at least as fast as the calculator itself[141].

A paper by Adams and Aizawa (2001) offers the opinion "Philosophers these days seem not to appreciate that isomorphism is a relatively weak relation". I wish to claim that, on the contrary, isomorphism is an exceedingly strong relation. Something physical which is fully, counterfactually (Chalmers, 1994; Chrisley, 1994), functionally isomorphic to a particular definition of a calculator is, in a good sense (quite the best sense, in fact) a calculator. I take it that I am with Dennett on this.

And I accept that RoboDennett can indeed perform a functionally isomorphic simulation of himself[142, 143]. As such (and again, I take it that I am with Dennett on this) what RoboDennett can do is generate a *bona fide* state of knowing what it is like. On this very strong functionalist account, RoboDennett has actually created an agent which knows what it is like. It is living in a virtual world, but it wouldn't necessarily know that this is the case (Chalmers, 2003b); it is up to the real RoboDennett to decide whether or not to make this information available to the simulation.

At this stage, though, the state of knowing what it is like is a state of the simulation, not a state of the simulating agent. Even on Dennett's account, to come to know what it

---

[141] Speed of simulation *is* important, here. We will look later at what heterophenomenology requires. If it turns out that there's any fundamental reason why RoboMary's simulation of herself is necessarily slower than the real thing, then we've got a behavioural distinction right there between a RoboMary who really knows what it is like and RoboMary who is just working out how to behave as if she knew what it is like, using a simulation.

[142] Indeed, I mean to allow that RoboDennett's simulation can meet *both* of the above two requirements. As far as accuracy goes, that the simulation can be sufficiently like RoboDennett for RoboDennett to know exactly what he *would* do, if exposed to colour, is *ex hypothesi*. As far as speed goes, I am not sure whether or not my description of RoboDennett entails that such simulation can be arbitrarily fast, and indeed this might depend on the use to which RoboDennett plans to put the simulation (basically, we can't allow this, if allowing it entails some contradiction) but my arguments won't hinge on this, either way.

[143] For reasons related to my preference for a non-reductive physicalism (which, I have argued, is no more nor less than normal science, see Section 2.2.3, and c.f. footnote 145), I am no longer sure that I wish to grant this unreservedly. Nevertheless, the point still holds that we *can* grant it (as Dennett would certainly wish to do) and still show that Dennett's line on the knowledge argument is incorrect.

is like, locked RoboMary has to do something above and beyond creating this simulation. She has to work out the relevant aspect of the state of the simulation (Dennett's state B), and then she has to *put herself into that state*. It is this step which RoboDennett cannot take. He can simulate himself as well as he likes[144], but that's it.

As I've said above (6.4.8), on my account of qualia, and from an engineering point of view, the state of knowing what it is like involves low level colour response circuitry becoming recruited such that it plays the right causal role in enabling noninferential space-of-reasons responses to (and as if to) colour[145, 146]. So now, we need to ask whether RoboDennett can make his low level colour processing visual 'circuitry' play the relevant causal role. If he cannot, we need to ask whether he can make *anything else* play the relevant causal role. If he can do neither of these things, then he simply will not be in the state of knowing what it is like, despite all his knowledge.

The first option above is unlocked RoboMary's route to coming to know what it is like: directly manipulating his early visual circuitry such that it is just as it would be if he was perceiving colour. We have already rejected it as cheating (Section 6.4.9), in quite a precise sense, and we need not consider it again. RoboDennett cannot do it.

What about trying the second option, of getting something else to play the relevant causal role? Again, RoboDennett can come tantalisingly close. He can't tamper with his actual colour categorisation system, but he can think very hard, and thereby bring into existence a perfectly good simulated colour categorisation system (indeed, one which is as it would be if he had seen colours). Now all he has to do is to put *that* simulation into

---

[144] In addition to the point made in the previous footnote, there would be problems if RoboDennett had to accurately simulate himself simulating himself in order to achieve his ends, since this might well entail an infinite chain of simulations. But once again, my arguments don't hinge on this point, and I'm prepared to allow that RoboDennett only needs to go one level deep in the simulation, and that he could unpick the differences in state due to the fact that he was running a simulation and the simulation wasn't, from those differences due to the fact that the simulation had experienced colour, and he hadn't.

[145] In order to situate this point more clearly in the context of the overall thesis, I should emphasize that this kind of 'low level circuitry' is important precisely because it enables whole-agent but sub-rational 'abilities' (or 'sensitivities'); the kind of abilities which come together to constitute the mental, but which are not, in and of themselves, mental (see the Appendix).

[146] The present arguments are given in terms of a particular functional analysis of knowing what it is like, but I believe Dennett would be wrong about RoboMary for the reasons expressed in Sections 6.4.7-6.4.10 on *any* functional account. I do not have an argument to establish that the conclusions of Section 6.4.11 follow if this and similar accounts are rejected.

the right causal relationship with those parts of his brain which enable his propositional reasoning abilities. Again, RoboDennett can do everything except the last step.

The ability to think very hard requires that an agent have very advanced, reason respecting transitions between its many and various thoughts. As we've mentioned, it also requires that there be *some* grounding of those thoughts in perception (not the particular sensory grounding which Mary doesn't yet have, but some grounding). There is no additional requirement that the agent be able to re-engineer, at will, the mechanisms governing all these reason respecting transitions, and this is what RoboDennett would have to do in order to use his simulated V4 to put himself into the functional state of knowing what it is like. On the present account, you know what it is like to see red only when you possess the ability to exercise the perceptually grounded concept that we might gloss as *'red_as_experienced'*. That concept exists only when the relevant linkage between low and high level brain circuitry – or something functionally isomorphic to it – has been created. To get this grounding other than by low-level stimulation of the kind which normally engenders colour experience, an agent would need to re-engineer its cognitive architecture using abilities which go beyond those *required* by the knowledge argument. Lacking this low level grounding, RoboDennett simply wouldn't have this grounded concept – with its concomitant behavioural and affective results – even though he knows *exactly* what these results would be, if he did have the grounded concept in question.

If RoboDennett would not 'know what it is like' to see in colour, even while he runs all these incredibly complicated simulations, we are entitled to ask what it *would* be like for him to run them. I submit that it would be like nothing so much as it would be like thinking very hard, with the concomitant 'intentional objects' such as inner speech and (non-coloured!) 'imagery', deriving from the sensorily grounded concepts which RoboDennett does have. As we have said, the result of all that thinking very hard would be that RoboDennett *would* know exactly what he should say and how he would react if he had seen colour. So now we have to address one final question[147]: why can't RoboDennett simply speak and react as he knows he should?

[147] Dennett was kind enough to press on me in person the fact that I hadn't properly addressed this final issue, in discussion of a conference presentation of an earlier draft of this paper.

6.4.11 What Heterophenomenology Requires

Dennett has frequently, eloquently and correctly argued that a difference that makes no difference *is* no difference (Dennett, 1991; Dennett, 1995; Dennett, 2004).

Take Dennett's position on philosophical zombies, for instance. A zombie is a creature which responds to any stimulus which experimenters present to it in exactly the way we would. Thus a zombie may well decide to stand there all day saying things like "Of course I have qualia! Why won't you believe me, dammit?", not just in the manner of an over-complex lookup table, but in *all* the same ways and on all the same occasions we would, tested and untested.

My gut reaction is that Dennett is quite right, that the correct response is to believe the zombie. Of course it has qualia[148]. To think otherwise is to make a fundamental mistake about the nature of introspection, a mistake which leaves each of us as the proud owners of our own epiphenomenal qualia. Of course, this is not (just) a gut reaction, much of this thesis has been a detailed argument in defence of it (see especially Chapters 2 and 3).

But we do need to make very sure that RoboDennett is not an unintended zombie. To sustain the claim that RoboDennett does not know what it is like, we need to demonstrate that he *cannot* behave exactly like a creature which does know what it is like.

I believe we can demonstrate this by first noting that whole-system level behaviour does not consist simply in verbal (or other types of: c.f. Marcel, 1993; Cowey and Stoerig, 1995) report. There are additionally many things that we, as agents, do, over which we have no conscious, voluntary control. We sneeze in response to dust; we blink to protect our eyes, and duck to protect our bodies from looming stimuli; we have certain innate, low level reactions to sound and, the case in point, to colour (Humphrey and Keeble, 1978).

If it is possible to build an agent who knows as much as Mary, but with our kind of hierarchical architecture, then these behavioural differences would remain. The very simplest example is speed of response: non-consciously mediated responses are simply faster (Marcel, 1993; Merikle, Smilek and Eastwood, 2001) than consciously mediated responses. Because of this, however much RoboDennett knows about how he should

[148] To more accurately reflect Dennett's position (though not mine), I should say: 'of course it is exactly as justified in claiming to have qualia as we are.'

have reacted to any given coloured stimulus which he sees, he will be too late to *actually* react as fast as if the reaction had genuinely been mediated by lower level processes. This is a *bona fide* behavioural difference, and one which RoboDennett cannot overcome.

There are also behavioural differences in kind, not just in speed, of response. Take the example of the heightened state of alertness in rhesus monkeys in response to red light reported by Humphrey and Keeble (1978). This change in behavioural pattern is mediated by an extremely complex set of biochemical changes, one which *we* very probably cannot create by any chain of conscious thought[149]; crucially, though, whether or not we actually can do this, it is entirely reasonable to suggest that there is no logical or physical *entailment* from the ability to understand what such changes consist in, to the ability to initiate such changes by any act of conscious will. Again, therefore, RoboDennett would lack these abilities, and simply would not be able to *behave* like a creature who had undergone the low-level changes which would occur in him after exposure to colour.

These low level abilities are a crucial part of what Mary gains, when she learns what it is like. She is said to know what it is like precisely because her more abstract concept, *'red_as_experienced'*, is partially constituted by the very systems which mediate faster, less abstract responses to red. A creature which really knows what it is like must really behave as if its low level systems have been exposed to colour, and it must also reason about colour, as experienced, in a way which is supported by those low level systems (with consequent two-way effects, from reasoning to low level responses and *vice versa*).

All of this RoboDennett would lack, despite his perfect knowledge of what he lacks. This will result in personal level behavioural differences, which he cannot overcome, between RoboDennett and an agent which does know what it is like.

Therefore, knowing as much as Mary does – knowing *exactly* what these low level behavioural differences consist in – does not entail the ability to behave differently, in this way, at will (it is *compatible with* such an ability, as in RoboMary, but it *does not* require it), thus RoboDennett, who can only do what he *must* be able to do in virtue of

[149] That is, assuming that something more or less analogous happens in us; or *mutatis mutandi*, if needed, to an example where something similar does happen in us.

his knowledge, does not know what it is like, *even on a strictly heterophenomenological account*.

6.5 RoboDennett and Ineffability

I have argued over the course of Section 6.4 (extracted from Beaton, 2005) that accepting that Jackson's Mary learns something on her release is no threat to physicalism. This, of course, is not an original claim in its own right. The contribution in the above is the response to Dennett's most recent work on the knowledge argument, for Dennett still argues that there is such a threat.

Of course, it has seemed to many over the years that if Mary learns something then there is indeed a threat to physicalism. However, many physicalists have offered compelling arguments against the existence of any such threat, *even if Mary does learn something*[150]. Recently, Jackson himself has joined the camp of those who accept that there is no threat to physicalism in Mary's learning something. As least as amongst those who argue *for* physicalism, Dennett seems to be ploughing a lone furrow on this argument. Now it is unwise to write off Dennett's lone furrows. They tend to be at worst well argued and informative, and at best – and often – correct despite the nay-sayers. In this instance, however, I believe I have offered strong arguments for the former outcome.

It can seem at first that Dennett *has* to find a threat in Mary's knowledge, for Dennett is the original heterophenomenologist and, according to heterophenomenology, the *only* data relevant to what it is like is what we say and do (Dennett, 1991). Surely, then, a heterophenomenologist has to believe that there is no:

> "distinction … between knowing "*what one would say and how one would react*" and knowing "what it is like"" (Dennett, 2005b footnote 3).

Not so. Knowing everything about what one would say and how one would react is *knowing what 'knowing what it is like' consists in*; whereas being in a position to actually react in that way is *knowing what it is like*. For all the reasons set out in Section 6.4, neither of these two states entails the other.

As such there is, on this account, something ineffable about qualia, for *you cannot put into words 'what it is like'*: no description, however extensive and careful, can be

[150] Knowing something is just gaining an ability, according to Lewis (1983); (therefore?) there is an equivocation on "knows" in the original argument, according to Churchland (1989).

sufficient to make someone who understands that description know 'what it is like' (merely in virtue of understanding the description).

This is a *bona fide* kind of ineffability: a case of inexpressibility in words. But, of course, it is only a limited ineffability. Even though no expression in words (however well expressed, and then however well understood) can be sufficient to let the recipient know what it is like, nevertheless a theorist can, on the present account, put into words exactly what knowing what it is like *consists in*[151].

6.6 Intrinsicness and Privacy

Summarising the previous sections briefly, I have argued that our knowledge of our own mental states is *infallible* and *incorrigible*, or rather, that we can make good sense of the claim that it is *in the nature* of such self-knowledge to be infallible and incorrigible (even though such knowledge can and does go wrong in real agents). For such states are *defined* by their rational role, and an agent cannot be *rational and wrong* in self-ascription of them. Equally, I have argued that there is some good sense to be made of the claim that qualia are *ineffable*, for you really cannot put into words 'what it is like', even though you can put into words what 'knowing what it is like' consists in.

What about intrinsicness and privacy? Clearly, in one sense, I have to deny these outright. Much of this thesis has been devoted to arguing that no part of the introspectible mind could be an intrinsic property, and that no part of mind is private in any strong sense (i.e. that 'the mental' and 'the public mental' are co-extensive). I have no intention of going back on that now. Instead, I simply wish to argue that much which the alleged intrinsicness and privacy of qualia (in particular, and of the mental in general) was meant to account for, can be accounted for on this present account, using only non-intrinsic, public mental properties.

---

[151] Since I would now prefer to endorse a non-reductive physicalism, this expression is arguably slightly inaccurate. I should perhaps better say: you can capture *arbitrarily well* (but, arguably, never perfectly), in words, what is going on in a system which knows what it is like. If this is right, then I think there is a further aspect of 'ineffability' present in this non-reductive relation between *explanans* and *explanandum*. Much more could be said on this, which space and time preclude. But, once again, there would be no threat to normal science here, for it would be exactly as appropriate (or otherwise) to argue that *all* scientific explanation is non-reductive, in the same sense.

6.6.1 Privacy

Firstly, privacy. Although I have made every effort to argue that it is of the nature of the mental to be accessible, *at least counterfactually*, to empirical validation, this does not prevent at least a weak 'privacy': it remains true that I cannot know what you are thinking just by looking at you. Any number of thoughts may be compatible with your current physical appearance. This does not gainsay the claim that these very same mental states should be analysed as mutually interacting, whole-system functional states[152], which are canonically identified *purely* on the basis of whole system behaviour. So, mental states on this view are private enough that the view ought not to be considered subject to the anti-behaviourist joke: "that was wonderful for you darling, but how was it for me?". We can know what we think better than we can know what others think, for we can introspect, which is a *different* way of gaining self knowledge than self perception. We get that much privacy, but no more. *What* we introspect are public mental states, in the above sense, *defined* in terms of their at least counterfactual effects on behaviour – whether we realise this or not.

6.6.2 Intrinsicness

Secondly, intrinsicness. Once again, I accept, indeed demand, that both our public and our introspected mental lives feature no intrinsic, non-relational features. But it is still worth recalling what the intrinsic aspect of qualia was supposed to buy us. On all the accounts canvassed in Chapter 2, it was supposed to buy us the possibility of different qualitative feel, as between two agents whose publicly accessible, mental level behaviour is the same. This, we can't have. But it is interesting to note that Shoemaker, who certainly *does* want to allow the logical possibility of behaviourally undetectable spectrum inversion (Shoemaker, 1975; Shoemaker, 1994c; Shoemaker, 1994d), at one point expresses this desideratum thus:

> "The intuition that this is so finds expression in the inverted spectrum hypothesis – it seems intelligible to suppose that there are creatures who make all the color discriminations we make, and are capable of using color language just as we do, but who, in any given objective situation, are confronted with a very different phenomenal character than we would be in that same situation, and it is not credible that such creatures would be misperceiving the world." (Shoemaker, 1994d p.24)

[152] Again, by saying 'state' I do not mean to rule out a dynamical, process-oriented account.

It turns out that we *can* allow that what Shoemaker says "seems intelligible", is indeed coherent, and quite possible. This is exactly what the present account of qualia buys us. For having said what a subject is rationally related to, and how (perception, memory, imagination[153], etc.) we have still *left out* something else required to get a full space of reasons account – we have left out any information about how the subject is motivated to react, in that situation. This missing motivation should not be thought of as a further relation to content (desiring state of affairs *x*, fearing state of affairs *y*, etc.). Instead, it should be thought of on a broadly adverbialist account: state of affairs *x* (i.e. what is perceived, etc.) is presented desirably (or fearfully, or painfully, as it might be), where the *very same* state of affairs, with *no* difference in intentional contents (in *what* there seems to be, to the subject), can also be presented in some different way. Qualia are thus identified with this additional, subjective, motivational and associative aspect of the space of reasons, in the ways described in Chapter 5. Qualia are *not* identified as that which *causes* the motivation, in a given subject (as in traditional accounts of intrinsic qualia), but as the behaviourally detectable motivation itself.

As such, these qualia are not intrinsic, but they do buy something very like what the traditional property of intrinsicness was supposed to buy; they allow a situation in which two different subjects discriminate exactly the same things, and can agree on a language to describe what they both discriminate, and yet have different qualia. It just needs to be carefully understood that the situation just described admits of behavioural (to wit, associative and affective) differences between the two subjects. It is here that the present account locates these subjects' qualia.

---

[153] It might be objected that in all these states, we can be (or seem to be) related to things which do not – even could not – exist. The response is a response which can be perceived in detail in Evans (1982): no sense is to be made of any mental relation to some intentional state of affairs, unless sense can be made of the claim that the subject *knows what it would be* (in a practical, rather than theoretical sense) for that state of affairs to obtain. The same point would seem to be made, if only in brief outline, by Shoemaker (1994d p.26) where he states that we can make no sense of a subject's hallucinating a ghost, unless "we at least have some idea of what would *count* as someone veridically perceiving … a ghost". This is *not* exactly the same point as Evans', but surely it's not too great a step from there to propose that *the subject* must likewise have some practical idea of what it is for there to be a ghost, in order to hallucinate there being one.

6.7 Summary

The purpose of this chapter has been to attempt to show – using the analysis of qualia offered in the previous chapter – that it is possible to naturalise many of the problematic intuitions surrounding qualia. Specifically, I have looked at their alleged *ineffability*, *intrinsicness* and *privacy*, and at our alleged ability to know them *infallibly* and/or *incorrigibly*. To the extent that this chapter has succeeded in explaining (rather than explaining away, c.f. Sections 2.2.7-2.2.8) such intuitions, this bolsters the claim that the introspectible, subjective properties identified in the previous chapter are indeed qualia. This is so, I have suggested, even though Dennett is correct to argue that many of the attempts to *codify* these intuitions have amounted to definitions of properties which nothing real can have.

I used the analysis of introspection developed earlier in order to argue that the properties I have defined should quite naturally be known infallibly and incorrigibly, in a certain sense: that one cannot be *rational and wrong* in self-ascription of them.

I used an extended argument against Dennett's most recent position statement on the knowledge argument, in order to show that a certain kind of ineffability (an inability to put 'what it is like' into words) is to be *expected* within – indeed, is *entailed* by – even the strictest of functional or heterophenomenological approaches.

Finally, although qualia as I have analysed them are fundamentally public properties, I have argued that they are 'private' enough to avoid the most obvious objections to a behaviourist or neo-behaviourist account. Similarly, I would agree with Dennett (1988) and others (Strawson, 1997; Smolin, 2000) that little sense can be made of strong intrinsicness; that *all* properties are, in the end, relational. Nevertheless I have argued that the basic force behind the intrinsicness intuition for qualia is that 'my red' might not be the same as 'your red', even if we are both seeing, and can successfully agree that we are both seeing, the same red sample as the same red sample. In lines of thought developed throughout this thesis, and merely repeated briefly above, I have explained how this can be so, albeit in a behaviourally *detectable* way.

In this way, and despite the very different metaphysical role for qualia on the present account as opposed to many standard approaches (Section 5.6), I have tried to reclaim[154] qualia from Dennett's repeated attempts to quine them.

---

[154] And not to 'foster': "foster [after John Foster], v. To acclaim resolutely the existence of something chimerical or insignificant" (Dennett, 1988 n.6)!

7. Conclusion

7.1 Concluding Remarks

This thesis has argued that we, as a scientific and philosophical community, already have the materials at hand to show that a scientifically respectable naturalisation of qualia is possible. It should be clarified that the arguments given here have not involved any attempt to locate low-level physical features responsible for the presence of qualia. Rather, the aim has been to provide an analysis which clarifies the nature of the high level properties standing in need of explanation. That said, it has certainly been argued that the analysis of qualia given here is fully *compatible* with eventual scientific explanation of the presence of qualia in terms of the presence of lower-level features, in the manner of normal scientific explanation. It is argued that this is a feature which many (probably most) analyses of qualia have not shared. Furthermore, it has been argued that this naturalisability has not been achieved at the expense of qualia realism. It has been argued that the high-level analysis of qualia given here can explain (and not explain away) many problematic intuitions concerning qualia: that they are knowable *infallibly* and *incorrigibly*, that they are *ineffable* and that they are (in technically 'weak', but theoretically important senses) *intrinsic* and *private*.

Chapter 2 presented an account of scientific explanation which is ubiquitous in the physical sciences (and elsewhere) whenever the existence of some higher level property is taken to have been explained in terms of the existence of some lower level property. It was made clear why *strong phenomenal realism* (the alleged logical possibility of zombies, full-blown inverted spectra, etc.) rules out any such explanation of qualia. It was argued that the modern phenomenal concept strategy cannot show (as it attempts to) that such problematic claims are compatible with physicalism. Next, it was suggested that many current and historical analyses of conscious *perception* have smuggled in implicit (and non-naturalisable) *theoretical* claims about the nature of *introspection*. It was suggested that strong phenomenal realism is guilty of this same sin. Given all this, it was argued that our quest for qualia should be guided by our best independently plausible theories of introspection. A form of *moderate phenomenal realism* was proposed, in which qualia were defined as introspectible properties which can vary, even as between two subjects who are seeing the same part of the public world

as the same part of the public world, and who can agree that they are doing so, in a shared, public language. It was argued that such introspectible, subjective properties (if they exist) would be sufficient to naturalise the inverted spectrum *intuition* (although it was clarified that nothing in the definition of such properties requires that they be behaviourally undetectable). It was accepted that if no such properties could be found, there would be no qualia (at least in one important, central sense of the term).

Chapter 3 presented just such an independently plausible theory of introspection: the rationality model defended tersely by Sellars (1956) and in detail by Shoemaker (e.g. in the papers collected in Shoemaker, 1996). An apparent (but only apparent) disagreement between Shoemaker and Sellars was presented and resolved. This has some implications for the relation between the personal level account of introspection being given, and the subpersonal level accounts which we might also reasonably hope to give. The rationality model was defended against certain recent objections. A novel argument was presented to the effect that the rationality model has a better claim to be counted as *bona fide* introspection than does the quasi-perceptual model against which it is often pitted. It was argued that Shoemaker's arguments for the rationality model can be presented in a generalised form which shows that *any* property of a space of reasons as such can be known in introspection (at least in principle, by the right kind of agent).

Chapter 4 presented Shoemaker's own most recent account of qualia. It was argued that this account can only be made compatible with Shoemaker's account of introspection if we are prepared to pay certain very high costs. In particular, Shoemaker's account of qualia rules out a causal account of our self-knowledge of phenomenal states, at both the subpersonal and the personal levels of analysis. The issue at the personal level should be particularly troubling to Shoemaker, since a causal account of the relations between mental states is what he otherwise endorses. It is accepted that Shoemaker may be aware of (at least some aspects of) these high costs, though he doesn't analyse these issues in detail himself, but it is argued that one should be very unwilling to pay such costs. Nevertheless, certain attractive features of Shoemaker's current account are highlighted (to do with the complexities of the ways in which we can think about our qualia, and about the public properties which cause our qualia in us).

Chapter 5 presents the central analysis of this thesis, in which it is claimed that qualia can be identified with introspectible (on the analysis of Chapter 3) subjective (in the

sense of Chapter 2) properties of a space of reasons as such: in particular, with those affective and associative properties which *must* be specified, if we are to give a space of reasons description of a subject in sufficient detail to have explicitly characterised the subject's reasons as reasons for *action*. This account is developed in more detail for the cases of colour qualia and pain. In the case of colour qualia, the previously mentioned attractive features of Shoemaker's account are adopted (though with important modifications at the most fundamental levels of the analysis) in order to account for the complexities of the various ways in which we can know our own qualia. In the case of pains, related lines of argument lead to the conclusion that it is wrong to try to completely eliminate pains, *qua* objects of perception (except as a misleading *façon de parler*), as was done in traditional adverbialist analyses of pain. It is argued that pains can be and should be identified with (at least seeming) body parts presented painfully: i.e. such that the body parts themselves become the subject's most direct reason for aversive action. In such a case, whilst *the pain* (in one sense of the word) is the body part sensed painfully, nevertheless the *feel* of the pain (another sense of the word pain) is the introspectible modification of the subject's space of reasons, which is such that the body part becomes a reason for aversive action. It is argued that the account can cope with the different feels of pain (sharp, dull, searing, throbbing, and so on). Significant differences between this account and traditional adverbialism are made clear; it is also argued that this account goes significantly further than traditional adverbialism ever went, in analysing qualitative feel. It is further argued that the account amounts to a form of direct realism. Direct realism (as it is to be understood here) is briefly defended and clarified. It is argued that this analysis of qualia, within the framework of direct realism, is also novel; some reasons why this should be so are presented.

Chapter 6 brings together work from earlier chapters to argue that qualia are knowable *infallibly* and *incorrigibly* (in quite strong senses) and also that they are *private* and *intrinsic* (in weak, but not unimportant senses). A novel response to Dennett's recent work on the knowledge argument is also given. This response to Dennett's work is then extended to show that qualia are also *ineffable*, in a certain sense: it is to be expected, even on a strictly physicalist account, that you cannot fully express 'what it is like' in words (at least, not in such a way that someone who has fully understood those words will thereby 'know what it is like').

It is argued that the above work has thereby been sufficient to reclaim qualia from Dennett's repeated attempts to quine them. For, on the above analysis: there are qualia; they determine what it is like to have an experience; and finally, on many quite normal occasions we, as theoretically informed subjects, are veridically aware of our qualia, as and when we have them.

7.2 Future Work

Several issues have been raised within the context of this thesis which could profitably be investigated further.

In Chapter 5, I proposed that qualia are directly introspectible on the rationality model of introspection, given an at least practical understanding of what qualia are: the subjective effect which public properties have on me. In a manner inspired by Shoemaker's work I also proposed that we can learn to directly (i.e. non-inferentially) perceive public properties *as having* that qualitative effect on ourselves (whichever subjective effect it is, in our own case). However, I think more analysis is merited on the issue of whether (and if so, in what sense) qualia are knowable *at all*, independently of their being individuated in the mind of a subject by the public properties which have such effects. In developing the analysis of Chapter 5, I was at least initially thinking of qualia (*qua* in principle public behavioural effects) as being knowable as such, independently of such individuation. Now, I am not so sure that this is correct. That said, it is at least arguable that the analysis of Chapter 5 as it currently stands does not in fact presuppose any such independent or prior knowability; but the issue could certainly be explored in more explicit detail. All of this, of course, directly relates to the pre-theoretic issue which Shoemaker discusses, of whether and in what sense "the smell of a skunk" (Shoemaker, 1994d p.25) is perceived as being entirely 'out there' (not an aspect of my mind), for all that it is, or can be (as above), directly perceived as having a fundamentally subjective aspect.

At points throughout the thesis, I have said that I endorse conceptualism, but I have not had the space and time to say a great deal about why (though see the Appendix for *some* detail). Effectively, it all comes down to whether or not there can be a reason to say that feature *x* of the world is present to a subject, as under description *x*, separably from any reason we may have to say that the subject has some at least practical understanding of what it is for something to be *x*. If there can be no presentation of the world to a subject outside the grasp of such understanding then, I would claim, this

makes conceptualism (McDowell, 1994) correct and nonconceptualism (Peacocke, 1992) false. However, there are many very subtle issues in play here. In particular, I think it has not been emphasized enough in the literature the degree to which arguments *for* conceptualism (McDowell, 1994) are tied to arguments *against* foundationalism (Peacocke, 2003). The conceptualist cannot just claim that perception is entirely structured within the categories of a subject's practical-rational understanding, although this often taken to be all that is at issue. The conceptualist also has to find a way to argue that *perception itself is an exercise of practical rationality*, with no extra-rational input *except for the world itself*. If the conceptualist cannot successfully defend this claim then some of the most central elements of what the nonconceptualist was arguing for turn out to have been correct all along. Clearly, then, there is much work to be done in further exploring these issues.

The above should also make it clear why I think there is a direct link between conceptualism, in its most tenable form, and direct realism, in *its* most tenable form (and 'direct realism' is certainly a portmanteau term, covering many wildly varying and some very unattractive positions). As already noted, I think that the analysis of qualia presented here *is* a conceptualist and direct realist thesis. But more could be said on whether and why this particular analysis of qualia does require either or both of these two controversial theses. Equally, more could be said on the connections between conceptualism and direct realism; and considerably more could be said (and needs to be said) in defence of direct realism, given the bad press it currently has amongst the majority (at present) of those philosophers of mind who aim to take science seriously.

One of the most central claims of the present thesis has amounted to this: the 'hard problem' as traditionally conceived (Chalmers, 1996) does not exist, for there is no *separate* problem of accounting for qualitative feel, above and beyond the problem of accounting for practical-rational behaviour. It has also been a key burden of this thesis to argue that this claim need not in any way amount to a denial of the claim that we have an immediate, subjective acquaintance[155] with phenomenal feel in all of our conscious doings. If an analysis of mind along the lines presented in this thesis proves robust then (even if, in all probability, considerably further in the future than anything else mentioned in this 'Future Work' section) there is clearly work to be done in addressing what I believe is the genuine hard problem: that of understanding what it is

[155] To be understood in the sense discussed in Chapter 2, footnote 41 and in Section 5.6.

about our physical universe which makes mind (understood *non-reductively* as realm of practical-rational behaviour) possible. Nothing here claims to have addressed *that* problem. Equally though, and to avoid misunderstanding, I should clarify that I do not necessarily believe that we need to go beyond a subtle appreciation of *current* physics, in order to start addressing such issues.

Appendix – Noë on Experience

Abstract

I review Noë's recent causal account of perception. I offer a formalisation of Noë's account, of a type which Noë himself gives for the old account which he argues against, but never gives for his own proposed replacement. In passing, I note that a worry which Noë himself offers, as to whether the terminology of his account is correct for the case of touch, can be fully dismissed. Finally, although I believe that Noë's account is a major step forward, I argue that it suffers from a notable flaw, in its own terms. Noë presupposes that there is a univocal sense in which we are related to what he calls *factual content* and to what he calls *perspectival content*. I argue that there are analytic reasons for believing that the relevant relationships in the two cases cannot be the same. I argue that Noë's account requires some small amount of sympathetic modification to allow for this issue, and I present the relevant modifications. I argue that the account gains something and loses nothing in the modification.

A.1 The Flawed, Gricean Theory

Noë (2003) credits to Grice (1961) the following causal analysis of perception.

A subject S sees an object O as being some way F if and only if:

- *S has an experience E, as of O being F*
- *O is F*
- *O's being F is causally responsible for the experience E*

Noë also credits to Grice (in Grice's later thought) and to others, the realisation that this theory cannot be complete. Take the example (Noë, 2003 p.93) of a manipulative neurosurgeon, who somehow (for instance, by direct stimulation of sensory cortex) causes an experience in S, which accurately reflects the way some object is, and where the neurosurgeon causes the experience to be that way because the object is that way. It is supposed this there is some sense in which this example is 'obviously' not true perception. If so, it is a counterexample to the Gricean theory.

Noë (2003 pp.98-99) also gives the example of 'Chris the amazing human hearing aid'[156]. Chris is a perfect mimic, and she (sic) is supposed to be able to repeat, into your ears, exactly the sounds you would hear if only… Actually, Noë's description of the

[156] Which Noë credits to David Sanford (unpublished).

example doesn't exactly specify "if only…" what. That is, Noë never clearly specifies why you cannot hear the sounds anyway, without Chris' aid. So perhaps the right formulation is, "if only Chris wasn't in the way, speaking into your ears". But, to make the example closely match the purposes of Noë's argument, let's specify that there is some kind of aural shielding around you, such that you would not hear how things are, if it were not for Chris the amazing mimic (re)creating the relevant sounds for you. Again, this situation is supposed not to qualify as genuine perception of the distal sounds in question, in some sense or other, even though it meets all the conditions of the unmodified causal theory.

Now, both these examples strike me as cases where there arguably is, and arguably isn't, perception of the relevant, distal objects. However, at least in the case of the manipulative neurosurgeon, it seems on first reading as if Noë is simply presupposing the plausibility of the above, as a counter-example to the simple causal theory. In fact, things aren't as simple as that: Noë eventually provides all the materials needed to show why our intuitions might (indeed, should) be mixed, about both of the above thought experiments. It turns out that both examples, when specified in no more detail than the above, can have further conditions added (with no modifications) in such a way that they either are or are not *bona fide* perception, on Noë's revised account.

A.2 The Project of Analysis

In the end, and as the title of his paper suggests, Noë's goal is to unravel a very central puzzle of perception: he aims to give a correct causal theory of it. Nevertheless, he commences with an admirable degree of modesty. He observes that:

> "it's doubtful that there has ever been an analysis (that is, a breakdown into necessary and sufficient conditions) of any philosophically interesting concept" (Noë, 2003 p.94).

This is certainly not by way of attempting to claim that he will be the first to achieve this goal. However, if that is not what Noë means to claim (and it is not), it might be unclear what he thinks he *has* achieved.

We can start to get clear about this by noting that Noë states that "the causal theory is obviously right in certain ways, and it is obviously wrong in others" (p.94). And that "it would be worthwhile to explain why this is so, even if we reject the project of analysis" (p.94).

Unpacking all the above points (in the light of the rest of his paper) Noë is saying something like the following: there almost certainly remain situations where our pre-

theoretic usages of 'perception', 'perceive', etc. outrun *any* causal analysis; nevertheless, the bare Gricean analysis given above seems to be wrong, even in its own terms. What is at issue, for Noë, is whether *any* analysis along Gricean (i.e. causal) lines is doomed to fail, in virtue of being unsuited to truly lock on to anything of interest in the area[157], or whether there is (as I believe Noë successfully demonstrates) some improved causal analysis which locks onto something central in the area (something which we might, with good grounds, call 'perception', even if this does involve *some* degree of relabelling with respect to pre-theoretic usage).

A.3 Noë's New Account

Noë also states, early on, that "[t]he problem with the causal theory is not that it fails to articulate with sufficient detail the right kind of causal relation" (p.94). This might also lead to some confusion since, if the reading of Noë I offer here is correct, the problem with the pre-existing, Gricean, causal theory is *precisely* that it fails to articulate the right kind of causal relation. The confusion (if any) is quickly removed once one realises that Noë's point is that Grice's account suffers not *simply* from a lack of detail (as regards that causal relation which it *does* consider), but rather, that there is a crucial further causal relationship which the Gricean causal analysis ignores entirely.

Noë's revised analysis is actually quite simple and natural (perhaps a good sign, in itself). He suggests that traditional (i.e. Gricean) causal analyses of perception have failed because they have attempted to analyse only an impoverished notion of the content of perceptual experience. The impoverished notion in question is that which takes perception to represent only *how things are*. The richer conception which Noë advocates holds that perception represents both 1) how things are, and 2) the observer's relation to how things are.

Noë's suggestion, then, is that previous causal analyses have failed to analyse perception, not by fault of being causal, but by fault of attempting to account only for the *factual*, and not the *perspectival* content of perception. Noë claims that any causal analysis of what is necessary and sufficient for perceptual presentation of how things are (the factual content) will continually fail to meet our intuitions concerning perception as such, precisely in virtue of its failing to be an analysis of what is necessary and sufficient for *both* perceptual presentation of how things are *and*

[157] As, for instance, Snowdon (1980-81) has argued.

perceptual presentation of the observer's relation to how things are (the perspectival content).

Let's look in more detail at the *perspectival* aspect of perceptual content, which Noë claims has been ignored, up to now, by those attempting causal analyses of perception. Noë states that:

> "we experience not only how things are, but also how they look from here. We experience that the plate is round and that it looks elliptical from here. Its elliptical look from here is a genuine property of the plate – we see the shape and we see the perspectival shape from here – but it is also a relational property, one that depends on where 'here' is." (p.95)

And further:

> "it is hard to understand how one could keep track of how things are if one were not also capable of keeping track of the ways in which one's perceptual experience depends on what one does. … [I]t seems likely that our practical grasp on the way [the perspectival shape of the plate changes] as we move is precisely the way we succeed in experiencing its roundness."

I think there is more to say than what Noë offers here (or elsewhere as far as I understand him) about the sense in which perspectival content is an aspect of *what* the subject perceives – an aspect of the "representational content of experience" (p.95), in the way in which Noë apparently intends this claim, i.e. that such things are *present for the subject*. I will explain what I mean in Section A.6 below. Nevertheless, I would fully endorse Noë's claim that:

> "To be a perceiver … you must understand, implicitly, that your perceptual content varies as things around you change, and that it varies in different ways as you move in relation to things around you." (p.97)

Perhaps surprisingly, Noë never explicitly spells out his revised causal theory in the same way in which he spells out Grice's theory. I will attempt to give such an explicit formulation here. As I understand it, Noë's revised theory can be expressed in the following formal claim.

A subject S sees an object O as being some way F if and only if:

- *S has an experience E, which is as of O being F, and is also as if S were in some perspectival relation R to O's F-ness*
- *S is in the perspectival relation R to O's F-ness (which entails: O is F)*
- *S's being in the perspectival relation R to O's F-ness is causally responsible for the experience E*

It seems to me that Noë is very much on to something here. For there is certainly a class of cases wherein we would like to get to the bottom of our various intuitions as to why certain experiences, caused in certain ways, are or are not *bona fide* cases of seeing (or of perceiving, more generally); cases such as those Noë discusses. I believe that Noë correctly draws our attention to the fact that previous attempts to analyse such cases, within a broadly causal framework, have proceeded on the assumption that the content of perception requiring analysis was only that which Noë calls *factual*.

A.4 An Analysis of the Counterexamples to Grice's Theory

Take the hypothetical neurosurgical example above. In the case *as specified*, it seems reasonable to presume that *if* the subject moves her eyes or head (or even her whole bodily location) then her experience will *not* track her actual perspectival relation to the objects of which she is being given experiences. In that case, there seems to be good reason to say that there is a valid sense in which this is not *bona fide* perception. I think Noë has put his finger on exactly what that good reason is.

On the other hand, we can vary our intuitions about this case in entirely the opposite direction, without contravening anything which was said, initially, in describing the example. Imagine, now, a subject in whom the relevant experiences are being generated in such a way that they *do* track not just what is there, but also the subject's relation to what is there. Is there any remaining reason to claim that this case is not (prosthetically assisted, but actual for all that) perception? It would seem not. And notions such as prosthetic perception are no longer mere idle speculations. If there *is* a sense in which someone who perceives prosthetically (as just described) is not truly perceiving, then so be it. There also seems to be *a* very good sense (the sense which Noë is aiming to clarify) in which the right kind of prosthetic perception is indeed an entirely valid, though entirely non-standard, form of perception.

In a similar vein, as Noë points out, when someone uses a hearing aid it is quite normal for us to say that they are (with the assistance of the hearing aid) hearing *the distal sounds*. Is there, then, any reason to deny that *the sounds themselves* are being heard, if Chris the amazing mimic can produce sounds which are faithful not just to what noise sources there are, but also to our perspectival relation to these things (such that what we hear gets louder when we move closer, quieter when we move further away, etc.)?

It might well be responded that, in this case, we hear Chris, herself, in the first instance, not the sounds. But there seem to be good reasons to question this response; reasons which support Noë's analysis. For again, it seems right to suggest that there is a valid sense in which we *do not* hear the hearing aid at all, when it is in our ear, and in use. And a good part of the reason why we might plausibly say this, of the hearing aid, is that we *don't have any potentially varying perspectival relation* to the sounds from the hearing aid itself, whilst using it. To the extent that we can't get such perspectival variation on Chris, then the fact that she is a person, not a machine, seems to make little difference: what we hear (*because* of her) are the sounds themselves, although we hear them in a non-standard way[158].

Indeed, Noë discusses (pp.93-94) one common way of attempting to strengthen the original, Gricean causal account, which involves requiring that *bona fide* perception must be caused "in the normal way". Noë points out two undesirable consequences of such a move. Firstly, it makes our account of what perception is (in itself, as it were) beholden to empirical discovery as to how perception works in our particular case. (Though perhaps, as Noë says, there are those who might think that such a result is something which philosophers should embrace, in the current intellectual climate.) Secondly, and decisively, the standard account strengthened in this way seems plain wrong, for precisely the reasons just canvassed. Imagine the case of the blind subject who has been given prosthetic vision, along the lines discussed earlier in this section. This is manifestly not vision caused "in the normal way", so it must be ruled out by the modified theory (along with the cases which we wanted to rule out), but there is surely *a* sense (this is the sense Noë is trying to elucidate) in which such a subject can, with the aid of the prosthesis, truly see the world.

A.5 The Perspectival Account and Touch

Noë himself appears to have some reservations as to whether his account applies across the board, to all the perceptual modalities. Specifically, he offers the concession that:

> "in the case of touch, the term 'perspectival' seems somewhat less appropriate" (p.95, n.2)

[158] Noë's point here is not that these various obscure, non-standard examples plausibly can reproduce our natural perceptual relation to things – but, rather, that it is only implausible that they are true perception *to the extent that* it remains implausible that they can do so (p.99, n.6).

This is a very natural worry. For surely there is no literal *perspective* in touch, in a univocal sense to that involved in the truly three-dimensional perception which can be given by vision, or even by hearing? In both of those cases, objects which are further away thereby come to *appear* closer together; they become harder to resolve, spatially. Nothing like this applies to touch, does it?

This is a very easy mistake to make, but it is a mistake. The *very same* rules of perspective apply to exploration of three-dimensional space by means of *touch alone*. Note, first, that there are well-attested cases where those blind from birth can paint in perspective (Kennedy and Juricevic, 2002; Kennedy and Juricevic, 2006 etc.). It turns out that this is neither inexplicable, nor fraudulent, nor need it be innate, nor even due to familiarity with what, we might wrongly suppose, would seem to the blind to be mere convention adopted by the sighted. This attested fact is straightforwardly (though non-obviously, except to very careful reflection) explicable by noting that touch *is* perspectival.

To understand for yourself how this is so, do not think about touching a two-dimensional surface ahead of you. Think, instead, about reaching out into three-dimensional space. Think, for instance, about the *movements* required to touch the nearby, accessible parts of 'receding' parallel lines. It turns out that the scare quotes around 'receding' are unnecessary, even if we are considering touch alone. If your two hands touch the nearby parts of the two parallel lines, the angle between your arms at your body is wide. If your two hands touch further parts of the same, *parallel* lines, then the angle made by your arms at your body is narrower. This is not something 'just like' perspective – it *is* perspective.

Perspective (and by this I mean to *include* the formal, mathematical treatment of the topic[159]) is all about the varying *directions* in space which are required to intercept near versus far things, whether this be for looking, for reaching, or for any other 'doing' in space. Exactly the same variations in these relative directions (and, for this reason, exactly the same formal mathematics of perspective) apply to vision, directional hearing, *and* touch.

It can additionally be noted (though this is not the central point) that if it still seems as if I haven't fully addressed *touch* (as opposed to reaching), there is yet more available evidence that Noë's account is the right one – even for touch. For it can also be noted

[159] In particular, the 1/z scaling of 'apparent', or projected, size with distance.

that one's perspectival relation to felt surface texture, say, is exactly as Noë would require: if one moves one's fingers in this or that way, over the surface, then one's contact with the surface texture varies in exactly the perspectival way which Noë's account requires.

A.6 The Problem For Noë's Account

It is clear that, to misquote, I come not to bury Noë's account, but to praise it. But I think there is a problem within the account which Noë would do well to resolve. This is meant – and I hope might be read by Noë – as a sympathetic amendment, a revision which helps to strengthen the account, within its own terms of reference.

My worry concerns the issue of whether a subject must "see the perspectival shape from here" (p.95), in order to "see the shape" (p.95), with a univocal sense of "see" in both cases. The easiest way to explain the substance of this worry is to use to terms of the debate concerning conceptual and nonconceptual content, in experience. In addressing this issue, I am more than happy to use an ability-based analysis of conceptual content (Evans, 1982 p.101) and, indeed, of content in general. This should fit very well with Noë's enactivist sympathies.

On an ability-based understanding of perceptual content, we work out the content of a subject's experience by working out what the subject is perceptually responsive to. On such a basis, we can only say that a subject has *conceptual* content (that is, brings an aspect of the world under the purview of a conceptual *ability*) to the extent that a subject responds[160] to the world as under some category, and where the responses in question are flexible, rational and, crucially, where the various conceptual responses can recombine arbitrarily.

As far as this latter point about arbitrary recombinability goes, I am trying to express, relatively informally, Evans' Generality Constraint on concept possession (Evans, 1982). According to this criterion for concept possession, a subject does not have the concept of red, merely in showing some categorial response to red. Additionally, their categorial response to red has to be recombinable, in arbitrary ways, with various other categorial responses to other aspects of the world.

---

[160] Actually, in any given situation, it is perfectly possible to have good evidence that a subject *is able* respond in a certain way, but doesn't choose to, or *would* respond in a certain way, if only tested slightly differently. For an ability-based account of perceptual content to be truly plausible, it has to additionally allow all the available evidence concerning these only counterfactual 'responses'.

Thus, if a subject responds reliably to a red ball, this is not yet enough to show that the subject has the concepts which might typically be labelled by the words 'red', and 'ball'. The subject must additionally show rational responsiveness to other red things, which are not balls, and to other balls, which are not red. Furthermore, these responses should be recombinable with a considerably broader range of flexible categorial responses, into significant, flexible 'occupation of' (i.e. behaviour within) a reasonable part of what is often called 'the space of reasons' (Sellars, 1956; McDowell, 1994; Hurley, 2003)[161].

So now, let us imagine a case where a subject shows the right kind of flexible, rational responsiveness for us to say that that subject has the concept[162] of 'circle' or of 'plate-shaped' or just of 'plate'. If a subject can respond visually to plates, say, with the kind of rational flexibility just described, then everyone on both sides of the conceptual/nonconceptual content debate would agree that the plate is present for the subject as conceptual content of their experience. Indeed, all sides, including Noë, would be happy to agree that, in this cases, the plate is unequivocally a part of Noë's *factual content* of experience[163].

We can also note that Noë is quite right: no subject could conceivably be shown to be successfully visually tracking plates, *qua* plates, across the range of relative motions of which plates are capable, unless that subject showed *some kind* of whole-agent-level sensitivity to what Noë calls perspectival content. For a subject to successfully demonstrate conceptual responsiveness to plates as such, across the range of cases in question, there has to be (*inter alia*) an experimentally verifiable ability to pick up on plates which are (according to the geometry of perspective) 'small from here', or 'large from here', or 'elliptical from here', or 'round from here'.

---

[161] This formulation, of course, defines concept possession not just in terms of Evans' Generality Constraint (nor would Evans himself have wished to define concept possession purely thus) but also adds other criteria, such as a requirement for evidence of rationality and flexibility in the exercise of any allegedly conceptual categorial abilities (c.f. McDowell, 1994).

[162] Concepts as just defined have *not* been defined in terms of possession of linguistic abilities (and at least arguably do not require the possession of such abilities). This is a standard usage of concept, on both sides of the conceptual/nonconceptual content debate (Peacocke, 2001 p.243; McDowell, 2007 p.347).

[163] I am not, at this point, presupposing that something *not* brought under concepts is not part of factual content; all I am saying here is that something which is brought under concepts, in this way, *is* part of factual content.

My central point, though, is that this latter type of responsiveness does not have to be conceptual. I agree that, for a subject to respond conceptually to plates, that subject *must* respond to smallness, largeness, ellipticalness and roundness of plates. But responsiveness to these latter properties in no way has to generalise, according to the Generality Constraint: the subject need show no flexible, rational responsiveness to ellipticalness as such, even though they must (logically must, if they possess the concept as described) be showing some kind of testable, verifiable, whole-agent sensitivity to perspectival-ellipticalness-from-here[164].

This point is not so pressing for a nonconceptualist. A nonconceptualist holds that there is some univocal sense of content, under which all aspects of the world responded to conceptually (as described above) and *at least some of the aspects of the world responded to nonconceptually* can together comprise content for (i.e. aspects of the world perceptually present to) a subject.

But my point should be particularly pressing for Noë, who endorses a brand of conceptualism (Noë, 2004 Ch.6), as do I. On a conceptualist account, *perceptual content (presentation of the world to a subject) consists in the active bringing of aspects of the world under a subject's concepts*[165]. Thus, for a conceptualist, in the case as described above, the plate is part of the content of the subject's experience, but the varying perspectival shape of the plate is *not* – or *need not* be. That is, a subject can be *aware of* a plate, as a plate, without needing to be at all *aware of* (though they must, in some sense, be *sensitive to*) the perspectival variations in the shape of the plate 'from here'.

Now, I think it might be easy to wonder whether I am not focussing on something which is no more than a mere 'slip of the pen' by Noë, in this particular presentation of his ideas. But I think that there is clear evidence that this is not so. For the implicit

---

[164] … -in-the-case-of-vision-of-plates!

[165] I do not wish to dismiss central nonconceptualist worries, such as those canvassed by Chrisley (1996). Certainly an infant has mental states wherein the abilities involved are far from fully conceptual. Indeed, I would agree with Chrisley's analysis under which the content in such a case might be at least partially determined by working out where the child's emerging understanding is going (would go, under normal conditions). What I still question is whether this notion of content is applicable, *other than to exactly the same extent that* the conceptual norms above are applicable. This is in much the same way in which belief and desire remain *defined* by their role in rationality, for all that real agents often show very significant irrationality about their beliefs and desires (Davidson, 1974; Dennett, 1987).

supposition that there is *some* univocal relation which we have *to things, and to the look of things*, runs throughout *Action in Perception* (Noë, 2004). To the extent that I can motivate a modification to Noë's causal analysis of perception, here, my suggestion is that a similar modification (or, at least, clarification) would sit well throughout the position set forth in that work.

Of course, we have already seen quotes which imply that Noë *to some extent* acknowledges this issue. This is presumably why he says that "To be a perceiver … you must understand, *implicitly*, that your perceptual content varies as things around you change" (p.97, emphasis added). But Noë only occasionally, and tacitly, acknowledges this complexity by use of words such as "implicitly" (this applies to both of his works referenced here). In the main, he is keen to emphasize that perspectival content is *represented* in experience (Noë, 2004 p.169 and passim); that is, that aspects of the world such as the 'looks' of things (e.g. Noë, 2004 p.168) are *present to the subject*. Apart from a tendency to use words such as "implicitly" in the perspectival but not the factual case there is no explicit acknowledgement of, nor any analysis of, the difference between these two cases. But there is a fundamental difference.

For the conceptualist, 'perspectival content' (in its required, minimal guise) cannot be considered true content at all; for the responses in question do not need to be conceptual, to complete Noë's story. Equally, it need not be supposed that we *see* both aspects of how things are (the factual and the perspectival) in any univocal sense. Instead, factual content is *bona fide* perceptual content (presentation of an aspect of the world to a subject) and perspectival content is not. We *see* the shape (in the case where we conceptually track that thing *as* being 'round') but we need not *see* the perspectival shape (in the same sense of see).

Even the nonconceptualist, who might well argue that both aspects of Noë's perceptual content *are* indeed content, in some univocal sense (and that we *always* see both the perspectival and nonperspectival shapes of the plate, in some univocal sense), should agree that there is *some* difference between the factual and perspectival cases – that is, should agree that in those cases where there *is* conceptual content and conceptual seeing of the objective shape, there *need not* be conceptual content nor conceptual seeing of the perspectival shape-from-here.

It might be wondered whether I can really be a conceptualist, given my endorsement of Noë's causal account (or, indeed, whether Noë himself can be), since that account makes these nonconceptual, perspectival relations between a subject and the objects of

their perception so central to the analysis of perception. But, I think, no conceptualist should ever have denied that perceivers must have nonconceptual behavioural *sensitivities* to aspects of their world. What is at issue is whether (as the nonconceptualist would claim), at least some things brought under the right kind of nonconceptual sensitivities are, thereby, perceptual *content* (aspects of the world, present for a subject), or whether (as the conceptualist should claim), such necessary nonconceptual sensitivities are, instead, active as *constitutive parts* of the conceptual abilities under which the world must be brought, in order for it to be present to a subject[166].

A.7 A Revised Account

We have seen, I think, that there is at least a lack of explicitness about this very important aspect of Noë's account; an aspect which ought to be particularly important to Noë, as a conceptualist, but which is not irrelevant even to the nonconceptualist. However, the revisions required to Noë's causal analysis of perception, to take account of this additional point, are not especially complex. Moreover, they can be expressed in a way which should be acceptable to both the conceptualist and the nonconceptualist, as follows:

A subject S sees an object O as being some way F if and only if:

- *S has an experience E, which is as of O being F, and where the subject is also (at least) nonconceptually sensitive (at least) as if to a perspectival relation R to O's F-ness*
- *S is in the perspectival relation R to O's F-ness (which entails: O is F)*
- *S's being in the perspectival relation R to O's F-ness is causally responsible for the experience E*

This formulation *allows* that 'seeing as' might be nonconceptual (although, as a conceptualist, I do not believe this is correct). What it clarifies is that the perspectival

[166] Much the same move can allow a conceptualist to be very sympathetic to, for instance, much of what Peacocke (1992) says about scenario content. *Of course* a perceiving subject must be nonconceptually sensitive to these aspects of the world; after having read Peacocke's analysis, that much need not be in doubt. What is in doubt, though, is whether 'scenario content' is really *content* (a bringing of the world into a subject's mind), or whether such nonconceptual abilities are 'merely' partially (if ineliminably) constitutive of those truly conceptual abilities, the exercise of which brings a world before a subject.

sensitivities required to support any given case of 'seeing as' certainly need not be conceptual, even if the 'seeing as' is conceptual.

Perhaps the revised account sounds overly technical, but I think there is a genuine non-technical reason to take it as a (sympathetic) improvement to Noë's account. For Noë says that we see shapes ('round', say) in and by seeing their perspectival shapes ('elliptical', say), but it is far from clear that we do. A naïve subject just sees a penny. Certainly, the subject must do so in and by being *sensitive* to the ellipticalness of the penny, 'from here'. But it is far from clear that the subject must *see* the penny by *seeing* the ellipticalness of the penny from here, at least not with any univocal sense of see.

Noë's account does indeed (at least to the present author's mind) unravel an important puzzle about perception, but it leaves this latter aspect puzzling. Here I have tried to show how to modify Noë's account to unravel this remaining puzzle, too.

A.8 Conclusion

I have argued that Noë's new causal account of perceptual experience has much to recommend it. I have suggested that it is entirely correct even in a domain (touch) wherein Noë himself worries that its validity might be limited. However, I have noted, there is a latent ambiguity in the account. Noë consistently states that we *see* shapes *and* that we *see* perspectival features of shapes (Noë, 2004); or, equally, that factual and perspectival content are both content, in some univocal sense (Noë, 2003). This, I argue, cannot be supported. There is an equivocation, here, and it is important that we get clear on what this equivocation is, if we are to truly get clear on what Noë's account comes to.

I have suggested that the correct tools to clarify this issue are those tools already used in the debate over whether perceptual content is conceptual or nonconceptual. Bringing these tools to bear is certainly relevant to Noë's work, since he himself has taken a clear stance on these issues (moreover, a stance which implies that he should be particularly worried by the points I raise).

The modified account which I offer may look as if it is overly technical. But, I have argued, the revised analysis can actually account for yet more of our pre-theoretic intuitions than can Noë's recent analysis, on which it is based. The revised version not only inherits from Noë his crucial insight that we *must* be sensitive to 'perspectival content' (the ellipticalness of the penny from here) in order to see, but also makes it clear (and in a way which matches pre-theoretic intuition) that this type of sensitivity is

not (or at least, need not be)[167] of the same type as the sensitivity which we have to the penny itself.

[167] For the conceptualist the correct formulation is: 'never is, in the most basic case of seeing as'.

References

Adams, F. and K. Aizawa (2001). The Bounds of Cognition. *Philosophical Psychology* 14(1): 43-64.

Alter, T. (1998). A Limited Defense of the Knowledge Argument. *Philosophical Studies* 90(1): 35-56.

Alter, T. and S. Walter (2006), *Phenomenal Concepts and Phenomenal Knowledge: New Essays on Consciousness and Physicalism* (New York: Oxford University Press).

Aydede, M. (2005/2008), 'Pain'. *The Stanford Encyclopedia of Philosophy (Winter 2008 Edition)*. E. N. Zalta, Ed.

Bauby, J.-D. (1997), *Le Scaphandre et le Papillon* (Paris: Editions Robert Laffont).

Beaton, M. (2005). What RoboDennett Still Doesn't Know. *Journal of Consciousness Studies* 12(12): 3-25.

Beaton, M. (in press), 'Qualia and Introspection'. *Journal of Consciousness Studies, Special Issue on 'Defining Consciousness'*. C. Nunn, Ed.

Braitenberg, V. (1984), *Vehicles: Experiments in Synthetic Psychology* (Cambridge, MA: MIT Press).

Byrne, A. and H. Logue (2009), 'Introduction'. *Disjunctivism Contemporary Readings*. A. Byrne and H. Logue, Eds. (Cambridge, MA: MIT Press).

Call, J. and M. Tomasello (2008). Does the chimpanzee have a theory of mind? 30 years later. *Trends in Cognitive Sciences* 12(5): pp.187-192.

Carroll, L. (1895). What the Tortoise said to Achilles. *Mind* 4(14): 278-280.

Carruthers, P. and B. Veillet (2007). The Phenomenal Concept Strategy. *Journal of Consciousness Studies* 14(9–10): 212–36.

Castañeda, H. and W. Sellars (1961-1962/2006), 'Correspondence between Hector Castañeda and Wilfrid Sellars on Philosophy of Mind (available at http://www.ditext.com/sellars/corr.html)'. A. Chrucky, Ed.**:** Quotes from Revised version: Aug. 4, 2006.

Chalmers, D. (1994). On Implementing a Computation. *Minds and Machines* 4(4).

Chalmers, D. (1995). Facing Up to the Problem of Consciousness. *Journal of Consciousness Studies* 2(3): 200-19.

Chalmers, D. (1996), *The Conscious Mind: In Search of a Fundamental Theory* (New York: Oxford University Press).

Chalmers, D. (2003a), 'The Content and Epistemology of Phenomenal Belief'. *Consciousness: New Philosophical Perspectives*. Q. Smith and A. Jokic, Eds. (Oxford: Oxford University Press).

Chalmers, D. (2003b). The Matrix as Metaphysics. *http://consc.net/papers/matrix.html* Accessed 7th July, 2005.

Chalmers, D. (2006), 'Phenomenal Concepts and the Explanatory Gap'. *Phenomenal Concepts and Phenomenal Knowledge: New Essays on Consciousness and Physicalism*. T. Alter and S. Walter, Eds. (Oxford: Oxford University Press).

Chalmers, D. and F. Jackson (2001). Conceptual Analysis and Reductive Explanation. *Philosophical Review* 110: 315-361.

Child, W. (1992). Vision and Experience: The Causal theory and the Disjunctive Conception. *The Philosophical Quarterly* 42(168): pp.297-316.

Chisholm, R. (1957), *Perceiving: A Philosophical Study* (Ithaca: Cornell University Press).

Chrisley, R. (1994). Why Everything Doesn't Realize Every Computation. *Minds and Machines* 4(4).

Chrisley, R. (1996), *Non-conceptual Psychological Explanation: Content and Computation* (DPhil Thesis: The University of Oxford).

Chrisley, R. (2008). Philosophical foundations of artificial consciousness. *Artificial Intelligence in Medicine* 44(2): pp.119-137.

Chrisley, R. (2009), 'Qualia: Realism without Cartesianism'. *ASSC 13*, Berlin, 5-8 June 2009.

Churchland, P. M. (1985). Reduction, qualia and the direct introspection of brain states. *Journal of Philosophy* 82: 8-28.

Churchland, P. M. (1989), 'Knowing Qualia: A Reply to Jackson'. *On The Contrary*. P. M. Churchland and P. S. Churchland, Eds. (Cambridge, MA: MIT Press)**:** 143-153.

Churchland, P. M. (1998), 'Postscript to Knowing Qualia'. *On the Contrary*. P. M. Churchland and P. S. Churchland, Eds. (Cambridge MA: MIT Press)**:** 153-157.

Churchland, P. M. and P. S. Churchland (1982), 'Functionalism, Qualia and Intentionality'. *Mind, Brain and Function*. J. I. Biro and R. W. Shahan, Eds. (Norman, Oklahoma: University of Oklahoma Press)**:** 121-145.

Churchland, P. M. and P. S. Churchland (1990), 'Intertheoretic Reduction: A Neuroscientist's Field Guide'. *Seminars in the Neurosciences***:** 249-256 (reprinted in Churchland and Churchland, 1998, pp.65-79).

Churchland, P. M. and P. S. Churchland (1998), *On the Contrary: Critical Essays, 1987-1997* (Cambridge, MA: MIT Press).

Clark, A. (2001), *Mindware* (New York: Oxford University Press).

Cowey, A. and P. Stoerig (1995). Blindsight in Monkeys. *Nature* 373: 247-249.

Crane, T. (2005/2008), 'The Problem of Perception'. *The Stanford Encyclopedia of Philosophy (Fall 2008 Edition)*. E. N. Zalta, Ed.

Davidson, D. (1974). On the Very Idea of a Conceptual Scheme. *Proceedings and Addresses of the American Philosophical Association* Vol. 47: pp.5-20.

De Jaegher, H. (forthcoming). Social understanding through direct perception? Yes, by interacting. *Consciousness and Cognition*.

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

Dennett, D. C. (1988), 'Quining Qualia'. *Consciousness in Modern Science*. A. Marcel and E. Bisiach, Eds. (Oxford: Oxford University Press).

Dennett, D. C. (1991), *Consciousness Explained* (Boston, MA: Little, Brown & Co.).

Dennett, D. C. (1994). Get Real. *Philosophical Topics* 22(1&2): 505-568.

Dennett, D. C. (1995). The Unimagined Preposterousness of Zombies: Commentary on T. Moody, O. Flanagan and T. Polger. *Journal of Consciousness Studies* 2(4): 322-326.

Dennett, D. C. (2004), 'Consciousness: How much is that in real Money?'. *Oxford Companion to the Mind, 2nd Edition.* R. L. Gregory, Ed.

Dennett, D. C. (2005a), *Sweet Dreams: Philosophical Obstacles to a Science of Consciousness* (Cambridge, MA: MIT Press).

Dennett, D. C. (2005b), 'What RoboMary Knows'. *Sweet Dreams: Philosophical Obstacles to a Science of Consciousness*. D. C. Dennett, Ed. (New York: Oxford University Press)**:** 103-129.

Dennett, D. C. (2006), 'What RoboMary Knows'. *Phenomenal Concepts and Phenomenal Knowledge: New Essays on Consciousness and Physicalism.* T. Alter and S. Walter, Eds. (New York: Oxford University Press).

Descartes, R. (1641). Meditations on First Philosophy.

Dienes, Z. (2004). Assumptions of Subjective Measures of Unconscious Mental States: Higher Order Thoughts and Bias. *Journal of Consciousness Studies* 11(9): pp.25-45.

Ducasse, C. J. (1942), 'Moore's Refutation of Idealism'. *The Philosophy of G. E. Moore*. P. Schilpp, Ed. (Chicago: Northwestern University Press).

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

Froese, T. (2009). Hume and the enactive approach to mind. *Phenomenology and the Cognitive Sciences* 8(1): pp.95-133.

Gertler, B. (2003/2008), 'Self-Knowledge'. *The Stanford Encyclopedia of Philosophy (Winter 2008 Edition)*. E. N. Zalta, Ed.

Graham, G. and T. Horgan (2000). Mary Mary, Quite Contrary. *Philosophical Studies* 99: 59-87.

Grice, H. P. (1961). The Causal Theory of Perception. *Proceedings of the Aristotelian Society* Suppl. Vol. 35: pp.121-152.

Hinton, J. M. (1973), *Experiences: An Inquiry Into Some Ambiguities* (Oxford: Oxford University Press).

Hume, D. (1739-1740/2000), 'A Treatise of Human Nature'. D. F. Norton and M. J. Norton, Eds. (New York, NY: Oxford University Press).

Humphrey, N. K. and G. R. Keeble (1978). Effects of red light and loud noise on the rate at which monkeys sample the sensory environment. *Perception* 7: 343-348.

Hurley, S. (2003). Animal Action in the Space of Reasons. *Mind and Language* 18(3): 231-256.

Jackson, F. (1977), *Perception: A Representative Theory* (Cambridge: Cambridge University Press).

Jackson, F. (1982). Epiphenomenal Qualia. *Philosophical Quarterly* 32(127): 127-136.

Jackson, F. (1986). What Mary Didn't Know. *Journal of Philosophy* 83(5): 291-295.

Jackson, F. (1998a), 'Preface'. *Mind, Method, and Conditionals* (London and New York: Routledge)**:** vii-viii.

Jackson, F. (1998b), 'Postscript on Qualia'. *Mind, Method, and Conditionals* (London and New York: Routledge)**:** 76-79.

Jackson, F. (2003), 'Mind and Illusion'. *Minds and Persons.* A. O'Hear, Ed. (Cambridge, UK: Cambridge University Press).

Jacobs, G. H. (1996). Primate photopigments and primate color vision. *Proceedings of the National Academy of Sciences* 93: pp.577-581.

Kant, I. (1996 (1781/1787)). *Critique of Pure Reason*. W. S. Pluhar, Ed. (Indianapolis: Hackett Publishing Company).

Kennedy, J. M. and I. Juricevic (2002), 'Optics and haptics: The picture'. *Multimodality of Human Communication: Theory, Problems and Applications*, University of Toronto, 3-5 May 2002.

Kennedy, J. M. and I. Juricevic (2006). Blind man draws using diminution in three dimensions. *Psychonomic Bulletin & Review* 13(3): pp.506-509.

Kind, A. (2003). Shoemaker, Self-Blindness and Moore's Paradox. *The Philosophical Quarterly* 53(210): 39-48.

Laureys, S., Ed. (2005), *The Boundaries of Consciousness: Neurobiology and Neuropathology*. Progress in Brain Research (Amsterdam: Elsevier).

Levine, J. (1983). Materialism and Qualia: The Explanatory Gap. *Pacific Philosophical Quarterly* 64: 354-61.

Levine, J. (2003), 'Experience and Representation'. *Consciousness: New Philosophical Perspectives*. Q. Smith and A. Jokic, Eds. (New York and Oxford: Oxford University Press).

Lewis, D. (1970). How to define theoretical terms. *Journal of Philosophy* 67: pp.427-446.

Lewis, D. (1980), 'Mad Pain and Martian Pain'. *Readings in the Philosophy of Psychology: Volume I*. N. Block, Ed. (Cambridge, MA: Harvard University Press).

Lewis, D. (1983), 'Postscript to "Mad Pain and Martian Pain"'. *Philosophical Papers: Volume I* (New York: Oxford University Press)**:** 130-2.

Lillard, A. (1998). Ethnopsychologies: Cultural variations in theories of mind. *Psychological Bulletin* 123(1): 3-32.

Loar, B. (1997), 'Phenomenal States (Revised Version)'. *The Nature of Consciousness*. N. Block, O. Flanagan and G. Güzeldere, Eds. (Cambridge: MIT Press)**:** 597–616.

Lycan, W. (2000/2008), 'Representational Theories of Consciousness'. *The Stanford Encyclopedia of Philosophy (Fall 2008 Edition)*. E. N. Zalta, Ed.

Marcel, A. J. (1993), 'Slippage in the unity of consciousness'. *Ciba Foundation Symposium No. 174 - Experimental and Theoretical Studies of Consciousness*. G. R. Bock and J. Marsh, Eds. (Chichester: John Wiley & Sons)**:** 168-186.

Martin, M. G. F. (2006), 'On Being Alienated'. *Perceptual Experience*. T. S. Gendler and J. Hawthorne, Eds. (Oxford: Oxford University Press)**:** pp.354-410.

McDowell, J. (1982). Criteria, Defeasibility and Knowledge. *Proceedings of the British Academy* 68: pp. 455-479.

McDowell, J. (1994), *Mind and World* (Cambridge, MA: Harvard University Press).

McDowell, J. (2007). What Myth? *Inquiry* 50(4): pp.338-351.

Merikle, P. M., D. Smilek and J. D. Eastwood (2001). Perception without awareness: perspectives from cognitive psychology. *Cognition* 79: 115-134.

Moore, G. E. (1903). The Refutation of Idealism. *Mind* 12.

Nagel, T. (1974). What Is It Like to Be a Bat? *Philosophical Review* LXXXIII(4): 435-450.

Nemirow, L. (1980). Review of Mortal Questions, by Thomas Nagel. *Philosophical Review* 89: 473-77.

Nisbett, R. and T. Wilson (1977). Telling more than we can know: Verbal reports on mental processes. *Psychological Review* 84: 231-259.

Noë, A. (2003). Causation and Perception: The Puzzle Unravelled. *Analysis* 63(2): pp.93-100.

Noë, A. (2004), *Action in Perception* (Cambridge, MA: MIT Press).

O'Regan, J. K. and A. Noë (2001). A sensorimotor account of vision and visual consciousness. *Behavioral and Brain Sciences* 24(5): 939-1011.

Papineau, D. (2002), *Thinking About Consciousness* (New York: Oxford University Press).

Peacocke, C. (1992), *A Study of Concepts* (Cambridge, MA: MIT Press).

Peacocke, C. (2001). Does perception have a nonconceptual content? *Journal of Philosophy* 98: 239-264.

Peacocke, C. (2003), *The Realm of Reason* (Oxford: Oxford University Press).

Perry, J. (2001), *Knowledge, Possibility, and Consciousness* (Cambridge, MA: MIT Press).

Pitcher, G. (1970). Pain Perception. *The Philosophical Review* 79(3): pp.368-393.
Price, H. and R. Corry (2007), *Causation, Physics, and the Constitution of Reality* (Oxford: Oxford University Press).
Puccetti, R. (1977). The Great C-Fiber Myth: A Critical Note. *Philosophy of Science* 44(2): pp.303-305.
Ramachandran, V. S. and S. Blakeslee (1998), *Phantoms in the Brain: Human Nature and the Architecture of the Mind* (New York: William Morrow).
Rosenthal, D. M. (1986). Two Concepts of Consciousness. *Philosophical Studies* 49(3): 329-359.
Rosenthal, D. M. (2002), 'Explaining Consciousness'. *Philosophy of Mind: Classical and Contemporary Readings*. D. Chalmers, Ed. (New York and London: Oxford University Press)**:** pp.406-421.
Sellars, W. (1956), 'Empiricism and the Philosophy of Mind'. *Minnesota Studies in the Philosophy of Science, Volume I: The Foundations of Science and the Concepts of Psychology and Psychoanalysis.* H. Feigl and M. Scriven, Eds. (Minneapolis, MN: University of Minnesota Press)**:** 253-329.
Sellars, W. (1963), *Science, Perception and Reality* (London: Routledge & Kegan Paul).
Sellars, W. (1975). The Adverbial Theory of the Objects of Perception. *Metaphilosophy* 6(2): 144-160.
Shoemaker, S. (1975). Functionalism and Qualia. *Philosophical Studies* 27: 291-315.
Shoemaker, S. (1982). The Inverted Spectrum. *Journal of Philosophy* LXXIX(7): 357-381.
Shoemaker, S. (1988). On Knowing One's Own Mind. *Philosophical Perspectives, 2, Epistemology*: 183-209 (reprinted in Shoemaker, 1996, pp.25-49).
Shoemaker, S. (1990). First Person Access. *Philosophical Perspectives, 4, Action Theory and Philosophy of Mind*: 187-214 (reprinted in Shoemaker, 1996, pp.50-73).
Shoemaker, S. (1991). Qualia and Consciousness. *Mind* 100: pp.507-524 (reprinted in Shoemaker, 1996, pp.121-140).
Shoemaker, S. (1994a). Self-Knowledge and 'Inner-Sense', Lecture I: The Object Perception Model. *Philosophy and Phenomenological Research* LIV(reprinted in Shoemaker, 1996, pp.201-223).
Shoemaker, S. (1994b). Self-Knowledge and 'Inner-Sense', Lecture II: The Broad Perceptual Model. *Philosophy and Phenomenological Research* LIV(reprinted in Shoemaker, 1996, pp.224-245).
Shoemaker, S. (1994c). Self-Knowledge and 'Inner-Sense', Lecture III: The Phenomenal Character of Experience. *Philosophy and Phenomenological Research* LIV(reprinted in Shoemaker, 1996, pp.246-268).
Shoemaker, S. (1994d). Phenomenal Character. *Noûs* 28(1): 21-38.
Shoemaker, S. (1995). Moore's Paradox and Self-Knowledge. *Philosophical Studies* 77: 211-228 (reprinted with revisions in Shoemaker, 1996, pp.74-93).
Shoemaker, S. (1996), *The First-Person Perspective and Other Essays* (Cambridge: Cambridge University Press).
Siegel, S. (2005/2008), 'The Contents of Perception'. *The Stanford Encyclopedia of Philosophy (Winter 2008 Edition).* E. N. Zalta, Ed.
Sloman, A. and R. Chrisley (2003). Virtual Machines and Consciousness. *Journal of Consciousness Studies* 10(4-5): pp.133-72.
Smith, A. D. (2002), *The Problem of Perception* (Cambridge, MA: Harvard University Press).

Smolin, L. (2000), *Three Roads to Quantum Gravity* (London: Weidenfeld and Nicolson).

Snowdon, P. (1980-81). Perception, Vision and Causation. *Proceedings of the Aristotelian Society* 81: 175-192.

Strawson, G. (1997). The Self. *Journal of Consciousness Studies* 4(5-6): pp.405-28.

Thompson, E. (2001), *Between Ourselves: Second-Person Issues in the Study of Consciousness* (Exeter: Imprint Academic).

Tomasello, M., J. Call and B. Hare (2003a). Chimpanzees understand psychological states – the question is which ones and to what extent. *Trends in Cognitive Sciences* 7(4): pp.153-156.

Tomasello, M., J. Call and B. Hare (2003b). Chimpanzees versus humans: it's not that simple. *Trends in Cognitive Sciences* 7(6): pp.239-240.

Tye, M. (1984). The Adverbial Approach to Visual Experience. *Philosophical Review* 93: 195-226.

Vimal, R. L. P. (in press), 'Meanings attributed to the term "consciousness": an overview'. *Journal of Consciousness Studies, Special Issue on 'Defining Consciousness'*. C. Nunn, Ed.

Wittgenstein, L. (1953/2001), *Philosophical Investigations (G. E. M. Anscombe (trans.))* (Oxford: Blackwell).

Woodward, J. (2001/2008), 'Causation and Manipulability'. *The Stanford Encyclopedia of Philosophy (Winter 2008 Edition)*. E. N. Zalta, Ed.