Emergence, Colour and the Knowledge Argument:

So What If Mary Didn't Know

Elizabeth Schier

Discipline of Philosophy
University of Adelaide

April 2007

ABSTRACT **4**

ACKNOWLEDGEMENTS **6**

CHAPTER ONE **7**

1.1 The Knowledge Argument **12**

1.2 The Standard Replies **14**
1.2.1 The No-Fact Reply 14
1.2.2 The Old-Fact Reply 17
1.2.3 The Status of the Argument 20

1.3 What Is the Problem with Mary? **22**
1.3.1 The General Strategy 23
1.3.2 Representation 26
1.3.3 Mandik's Account of Mary's Ignorance 32

1.4 What About Consciousness? **39**

1.5 Overview **43**

CHAPTER TWO **47**

2.1 The Psychophysical Data **53**
2.1.1 The Spectral Nature of the Surround 54
2.1.2 Space, Figure and Colour 58
2.1.3 The Challenge for a Representational Account of Colour Experience 62

2.2 A Complex Visual Model **66**
2.2.1 Achieving Constancy 66
2.2.2 Additional Empirical Considerations 73

2.3 Modelling the Visual World **79**
2.3.1 Molecular and Lexical Schemes of Representation 79
2.3.2 The Molecular Account of Vision 82

2.4 The Colour Vision Facts **88**

CHAPTER THREE **92**

3.1 Emergence **94**
3.1.1 The General Notion 94
3.1.2 Epistemic Notions of Emergence 97
3.1.3 Metaphysical Notions of Emergence 100

3.2 Mereological Inexplicability and Novelty in Science **106**
3.2.1 Water 106
3.2.2 Superconductivity 110
3.2.3 Oxidative Phosphorylation 113

3.3 Structural Emergence **116**
3.3.1 The Account 117

3.3.2 Multiple Realisability and Constitution Without Identity 124

3.4 The Emergence of the Colour Visual Facts 130

CHAPTER FOUR 133

4.1 The Epistemological Account 134
4.1.1 The Complexity of Visual Representations 135
4.1.2 Fragmented Representations 140
4.1.3 Resources, Content and Fragmentation 145
4.1.4 Knowing Structurally Emergent Phenomena 149

4.2 Structural Emergence and Materialism 151

4.3 Mary's Ignorance and Materialism 158

4.4 The Knowledge Intuition 167

CONCLUSION 172

REFERENCES 177

Abstract

One of the most problematic intuitions we have concerning consciousness is that the only way to know what an experience is like is to have the experience. For example it seems just obvious that the deaf person does not know what middle C sounds like. This intuition has been used to argue for dualism (Jackson 1982; 1986). The problem is that other physical things can be known in many different ways. It doesn't seem to matter what sorts of experiences you can have when it comes to knowing physics or chemistry. However, what sorts of experiences one can have seems crucial when it comes to knowing what an experience is like. In this thesis I argue that, contrary to initial appearances, materialism is not threatened by this "knowledge intuition". My general strategy is to develop an account of the facts that are represented in colour experience on which there are physical facts that we can represent only by having colour experiences.

I begin by developing an account of the represented object of colour experience that can do justice to the psychophysical data. It is argued that if we adopt a molecular scheme of representation, it is possible to understand the interdependence between the representation of the spectral, spatial and figural features of the world in a representational way.

Then I develop a "structural" account of emergence according to which a complex phenomenon is metaphysically distinct from its constituents because it has casual powers that are dependent on the physical organisation of the constituents. I argue that a structural account of emergence captures what science has already told us about the world and deals with the problems with causation that have plagued other accounts of emergence. I then argue that the facts that we know in colour experience are structurally emergent such that the complex state of affairs that is the constituents in arrangement is metaphysically distinct from the constituents.

Finally I argue that there are facts that we can know only via colour vision because it is only in vision that we have the resources to build a representation of the emergent phenomenon that is the parts in relation. However, because it is possible to know the constituents and the arrangements of a structurally emergent phenomenon without knowing the phenomenon itself, it is possible both to know that the phenomenon is physical and to obtain an explanation of it without actually knowing it.

I conclude that it is possible for there to be physical facts which we can know only via experience. The upshot is that materialism can be true regardless of whether or not the knowledge intuition is true. We will see that what is required to put consciousness on a par with other obviously physical phenomena is not a denial of our intuitions about consciousness, but rather a proper understanding of the metaphysics and epistemology of other physical phenomena.

Thesis Declaration

This work contains no material which has been accepted for the award of any other degree or diploma in any university or other tertiary institution and, to the best of my knowledge and belief, contains no material previously published or written by another person, except where due reference has been made in the text.

I give consent to this copy of my thesis being made available in the University Library.

The author acknowledges that copyright of published works contained within this thesis (as listed below) resides with the copyright holder(s) or those works.

Acknowledgements

Many thanks to:

My supervisors extraordinaire: Gerard O'Brien and Jon Opie for teaching me about representation, metaphysics and how to be a philosopher. This project would not have been possible without their expertise, support, critical guidance and ability to convince even the most stubborn person that they might actually be wrong.

My parents for their unwavering belief and editorial advice.

Glenn Carruthers for letting me talk at him in an attempt to work out what I was thinking, for editing numerous drafts and for always telling me that I could finish, even when I didn't believe him.

Louise Schier for not laughing at me too often despite my tendency to loose touch with the real world and forget what those, um, ah, err, word things are.

Members of the departments of philosophy at Adelaide and Flinders Universities, particularly: Wayne Christensen; Garret Cullity; Philip Gerrans; Rosemary Lowry; Greg O'Hair; and Nicole Vincent.

The material in Sections 2.1, 2.2 and 2.3 is adapted from:

Schier, E. (2007). "The Represented Object of Color Experience." *Philosophical Psychology* **20**(1): pp.1-27.

Chapter One

The Knowledge Argument

Consciousness is wondrous. It makes life worth living and keeps philosophers of mind in employment. It is at the centre of our mental lives and yet remains intolerably puzzling. The phenomenal quality, or "what it is like" of an experience, is how things appear in experience. It is the stinging sharpness of a pinprick, the sweet bitterness of dark chocolate, or the mindless ecstasy of victory. We go out of our way to seek out certain phenomenal experiences: the full-bodied sharpness of a good Australian Shiraz; the intoxicated relaxation of too many glasses of a good Australian Shiraz; or the hysterical laughter of a good joke. Other experiences are so unbearable that we try to avoid them at all cost: the throbbing pressure of a sinus infection; the bleak apathy of depression; or the desolate isolation of a broken heart.

Given how central consciousness is to our existence, you would expect that we would understand its place in the world. But, it is this very idea – that we have some intimate first-person contact with our experiences – which provides such an obstacle. To many it seems almost nonsensical to claim that the lump of grey matter in one's head can give rise to such a diversity of experiences. How can the electrical and chemical activity across a system of neurons feel like anger or

love? How can the electrical and chemical activity across a system of neurons feel like *anything*?

The extent of this bewilderment is so great that some have used it as the basis of their claim that consciousness is not physical (Chalmers 1996; Jackson 1982; 1986; Nagel, T. 1974; 1986; 1998; Searle 1992). Such dualists claim that consciousness is so completely different from the rest of the physical world that it must be a completely different, non-physical part of the world. Although there are many things about the mind that we currently do not understand, it is claimed that there is something about consciousness that makes it impossible, *in principle,* to explain. We supposedly have no idea how any physical thing could be phenomenal. As Nagel puts it 'the idea of how a mental and physical term can refer to the same thing is lacking' (Nagel, T. 1974, 447). The claim that consciousness is physical, or that experiences are brain states, seems to be as nonsensical as the claim that a number can have parents (Nagel, T. 1998).

This claim, that consciousness is necessarily puzzling, is supported by appealing to various pre-theoretical intuitions. Dualists develop "intuition pumps," arguments that are designed to evoke our pre-theoretical intuitions; in order to support their claim that "consciousness is physical" is nonsensical or inconceivable. One of the most famous – and as we will see problematic – of these intuition pumps is Frank Jackson's knowledge argument (1982; 1986).

The knowledge argument pumps the intuition that the only way to know what an experience is like is to have the experience. Call this the "knowledge intuition". This seems self-evident. It is just obvious that the blind person doesn't know what it is like to see red; that the deaf person doesn't know what it is like to hear middle C; or that the American does not know what it is like to taste Vegemite.

The problem is that it seems that if the knowledge intuition is true then it seems that phenomenal experiences are beyond the reach of physical science. How can the blind scientist find out what red looks like? As Smart (1987) suggests, when we adopt materialism we see the world apart from any particular perspective. In contrast, knowledge

of experiences seems to depend on your particular experiential perspective. A photometer can tell you the distribution of the photons being reflected from a surface, but it cannot capture what it is like to see red. A mass spectrometer can tell you the precise chemical composition of an eighteen year-old scotch, but it cannot capture what it tastes like. A precise characterization of all of the sound waves emanating from the orchestra will not tell you what the symphony sounds like.

Our scientific knowledge appears to be experience-independent. It doesn't matter whether you are a bat, Martian or human, you can know that lightning is electrical discharge. But, only those with vision can know how lightning appears to us in experience (Nagel, T. 1974). Nagel writes:

> If the facts of experience – facts about what it is like *for* the experiencing organism – are accessible from one point of view, then it is a mystery how the true character of experiences could be revealed in the physical operation of that organism. The latter is a domain of objective facts *par excellence* – the kind that can be understood from many points of view and by individuals with differing perceptual systems (Nagel, T. 1974, 442)[1]

It seems that the only way to know what an experience is like is to have the experience. In contrast, knowledge of other physical things appears to be independent of experience. It appears that phenomenal experiences are known in a different way to other physical things. Importantly, the most obvious reason as to why consciousness is known in a different way to other physical things is because it is not physical. As we will see in more detail below, the knowledge argument pumps the intuition that what it is like to have an experience is known

[1] Although Nagel often appears to be making metaphysical claims when talking about the objective and subjective, it is clear that he takes these to be primarily epistemic notions. He writes: 'Objectivity is a method of understanding. It is beliefs and attitudes that are objective in the primary sense. Only derivatively do we call objective the truths that can be arrived at in this way.' (Nagel, T. 1986, 4).

in a different way to other physical things in order to argue that what it is like is not physical.

Materialists respond to intuition pumps such as the knowledge argument by suggesting that puzzles can be solved and that there is no in principle reason why consciousness cannot be physical. They point out that our commonsense intuitions often turn out to be wrong (Churchland, P.M. 1995; Churchland, P.S. 1996; Dennett 1991). For example, at one point in time it seemed “just obvious” to various people that the earth was flat; that the sun orbited the earth; and that life was not physical. Yet as our understanding of the world developed, the “just obvious” became just false. Many materialists suggest that the same thing may happen with consciousness. With a deeper understanding of the brain, perhaps the just obvious “facts” that consciousness cannot be physical and that the knowledge intuition is incompatible with materialism will be shown to be false.

There is unquestionably something right about this type of reply. As Paul Churchland has pointed out, 'the fact that some thing, x, is mysterious to us is a fact about *us*, a regrettable fact about our current cognitive state. It is not a "telling" fact about x, a fact from which heavy-duty metaphysical conclusions might be drawn’ (1995, 191). The fact that we don’t understand how consciousness can be physical is a problem with *us,* not necessarily with consciousness. However, the debate cannot be resolved quite this effortlessly. As Churchland quite rightly goes on to say, ‘neither can substantive theoretical questions be decided just by citing some carefully chosen examples from the history of science ... We may rightly seek instruction from our intellectual past, but each new theoretical issue must ultimately be decided on its own merits' (Churchland, P.M. 1995, 191). An examination of the history of science *suggests* that we need to be careful drawing inferences from our confusion. Ultimately what will decide the issue is the nature of the phenomenon itself. It is only by examining the mysterious phenomenon from every angle that we can hope to either solve the problem, or establish more firmly that it is genuinely mysterious.

In an implicit recognition of this fact, materialists have made a considerable effort to show what is wrong with the knowledge argument. Yet the debate remains unresolved. In this chapter we will see that the problem with the standard replies to the knowledge argument is that although philosophers have found many different ways of responding to the argument, there is disagreement as to what precisely is wrong with the argument. As Van Gulick suggests, the diversity of materialistic responses is problematic because it

> might be taken to show that physicalists [materialists] have failed to find a clear answer to its challenge. If they cannot agree about where or how it goes wrong, can they be confident that it does so? (2004, 401).[2]

The mere multiplicity seems to be evidence of the inadequacies of the replies. It suggests that no one has found *the* reason why the knowledge argument fails.

The aim of this chapter is to try and home in on the remaining problem presented by the knowledge argument. I will argue that although the standard materialistic responses to the knowledge argument do secure materialism, they do this at the expense of avoiding dealing with *all* of the possibilities. In particular, they fail to deal with the possibility that there are facts that we can know only via experience. The problem is that we do not know enough about the mind/brain to know if the knowledge intuition is true. Given our uncertainty, the best strategy is to cover all options. I will argue that the only way to ensure the onus is shifted entirely on to the dualist is to demonstrate that facts that we can know only via experience can be physical.

In the first section I will examine the knowledge argument in more detail. In the second section I will examine the standard replies to the argument – all of which entail finding various ways to *deny* that there are facts that we can know only via experience. We will see that although these replies provide accounts that are compatible with

[2] Note that Van Gulick (2004) goes on to suggest that there may be more commonalities between the replies than there initially appears to be.

materialism, as a whole they are inadequate simply because they fail to deal with an obvious option; namely that there *are* facts that we can know only via experience. In the third section I will examine the work of others who have attempted to show that materialism can be true even if there are facts that we can know only via experience. We will see that although they make some important first moves, no one has developed a detailed account on which there are physical facts that we can know only via experience. In the fourth section I will examine what we are committed to regarding consciousness given that we have accepted that there are some facts that can be known only via experience. We will see that the knowledge argument actually has very little to do with consciousness. By the end of the chapter we will have determined what needs to be done in order to demonstrate that facts that we can know only via experience can be physical. In other words, we will home in on the tasks that need to be completed if the materialist is to have a comprehensive and decisive range of replies to the knowledge argument.

1.1 The Knowledge Argument

The knowledge argument is one of the most popular arguments for dualism. It is based in a thought experiment about Mary, a super neuroscientist. As Jackson puts it:

> Mary is confined to a black-and-white room, is educated through black-and-white books and through lectures replayed on black-and-white television. In this way she learns everything there is to know about the physical nature of the world. She knows all the physical facts about us and our environment, in a wide sense of "physical" which includes everything in *completed* physics, chemistry and neurophysiology … If physicalism is true, she knows all there is to know … It seems, however, that Mary does not know all there is to know.

> For when she is let out of the black-and-white room or given a colour television, she will learn what it is like to see red, say (Jackson 1986, 271)

Mary knows all of the scientific facts. In particular she knows everything science can tell her about what goes on in our brains when we see ripe tomatoes or look at the sky and use terms like "red" and "blue". Jackson (1986, 291) suggests that this complete scientific knowledge equates to complete physical knowledge. But, Mary has never had colour experiences. It *seems* that when she sees a ripe tomato for the first time, Mary would learn something, namely what it is like to see red. But, she supposedly had complete physical knowledge before her release. This suggests that the new knowledge she gains is not physical. It appears that there is something about consciousness that physical science can't tell us.

The basic structure of the knowledge argument is:

'1) Mary (before her release) knows everything physical there is to know about other people [and the world in general]

'2) Mary (before her release) does not know everything there is to know about other people (because she *learns* something about them on her release)

'Therefore,

'3) There are truths about other people (and herself) which escape the physicalist story' (Jackson 1986, 293).[3]

Mary has complete physical knowledge but she does not have complete phenomenal knowledge. The fact that what it is like is known in a different way to other physical things suggests that it is not physical.

[3] Note that one of the things that participants in the debate regarding Mary have focused on is the most precise way to articulate the knowledge argument (Nida-Rümelin 2002 provides a good review of this literature). Although it is important to precisely articulate the argument, the details of this debate are not germane to the current discussion. This is simply because I am interested in examining the knowledge intuition rather than any particular argument constructed around the intuition.

1.2 The Standard Replies

By far and away the most popular approach to the knowledge argument has been to analyse the new knowledge that Mary gains in such a way that it is compatible with materialism. In particular, the aim is to show that Mary does not gain knowledge of a new fact.[4] The key to these replies is to point out that just because Mary gains a new way of knowing does not mean that she learns about a new fact. What is needed, if the knowledge argument is to support any dualistic metaphysical conclusion, is not just that experience provides a special way of knowing, but also that it provides knowledge of something new. It is only if there are *facts* that we can know only via experience that the knowledge argument could present a problem for materialism. Most materialistic responses to the knowledge argument claim that although Mary may gain new knowledge, she does not gain knowledge of a new fact. They do this in two main ways: by denying that Mary gains factual knowledge; or by denying that Mary gains knowledge of a new fact.

1.2.1 The No-Fact Reply

One reason why the knowledge argument would fail to support dualism is if experiential knowledge were not factual knowledge. If Mary's new knowledge is not factual, then there can't be non-physical facts that she learns about upon release. This is the basic idea behind the "no-fact" reply. The difference between no-fact replies is the account of Mary's new non-factual knowledge that is provided. The standard version of the no-fact reply holds that Mary's new knowledge is an ability (Churchland, P. S. 2002; Lewis 1990; Nemirow 1990), while Conee (1994) argues that Mary's new knowledge is knowledge by acquaintance. What these responses share is the idea that although Mary gains new knowledge, this new knowledge is not factual.

[4] I take a fact to be a state of affairs; 'a way the world is' (Alter 1998, 42). For more on facts see chapter two.

Factual knowledge-that is often contrasted with the knowledge-how that underlies abilities. When I say that I know how to ride a bike, I seem to be talking about a different type of knowledge than when I claim to know that Canberra is the capital of Australia. Knowing how to ride a bike is a skill; it is something that you *do*. In contrast, knowing that Canberra is the capital of Australia is not a skill; it is knowledge of a fact. Know-how is not about some fact in the same way as knowledge-that. An ability is knowledge of how to *do* something, not knowledge *of* something. The idea behind the ability version of the no-fact reply is that Mary's new knowledge is an ability, and therefore does not provide her with knowledge of a new fact.

Consider for example Patricia Churchland's (2002) reply to the knowledge argument. She begins by pointing out that Mary has been prevented from having the brain events that are the experience of colour. The "neuronal theory" of colour vision, of which she has complete knowledge, has never been true of Mary herself. When she experiences colour for the first time, Mary now has the neuronal events that she has been studying. Churchland claims, however, that when she experiences colours for the first time, Mary is not learning *about* anything new. Rather she learns how to *do* something new. She gains a new perceptual skill: the ability to discriminate and recognise colours. The point is that

> to have the perceptual skill of discriminating and recognizing colours requires more than just knowledge of the theory of how our colour-discriminating system works: it requires that the theory also be *true* of oneself. It requires that *one actually possess a functioning instance of the neuronal system that the theory describes* (Churchland, P. S. 2002, 192).

The idea is that when Mary experiences colour and comes to possess a "functioning instance of the neuronal system that the theory describes", she is not learning any new facts. Rather she gains the ability to discriminate between colours. She gains the ability to *use* the neuronal system to distinguish between colours. In other words,

Churchland's proposal is that Mary gains the *ability* to see colours, not new factual knowledge.

Other proponents of the ability version of the old-fact reply analyse the ability Mary gains in different ways. Nemirow (1990) suggests that what Mary gains is the ability to *imagine* what red is like. Lewis (1990) suggests that knowledge of what it is like is the ability to *remember*, *recognise and imagine* the experience. However, they share with Churchland the idea that Mary's new knowledge is not factual, but an ability.

Like the ability version of the no-fact reply, the no-fact acquaintance reply holds that Mary's new knowledge is not factual. The difference between the replies is the analysis of Mary's new knowledge that is provided. Conee (1994) suggests that what Mary gains is not knowledge-that, nor knowledge-how, but knowledge by acquaintance:

> To have acquaintance knowledge is neither to know facts nor to possess abilities, but to stand in some sort of direct epistemic relation to someone or something. It is knowledge in the sense in which you know your colleagues, and in which you do not know a famous author you have read but never met (Alter 1998, 38).

The idea is that you may know all of the *facts* about a person or place, but still not know everything about them because you are not directly acquainted with them. For example, although you may know all of the facts about Brad Pitt or Daniel Dennett, unless you have actually met them, there seems to be a sense in which you do not know them. You are not acquainted with them. Conee suggests that when Mary has colour experiences she becomes acquainted with them; she enters into the maximally direct cognitive relation with them. This new acquaintance knowledge is not factual. What Mary becomes acquainted with is the look of the colour, its phenomenal *properties.* According to Conee, this knowledge of properties is not factual because factual knowledge is knowledge of truths, and a property is not a truth (Conee 1994, 140). The idea is that although pre-release Mary lacks knowledge, because the knowledge she lacks is

acquaintance knowledge, there is no new fact that she learns upon release.[5]

What both the ability and the acquaintance reply have in common is the idea that Mary's new knowledge is not factual. Analysing Mary's new knowledge as non-factual is important, because it means that her pre-release knowledge was factually complete. The idea is that there is no need to propose the existence of non-physical facts that are the objects of Mary's new knowledge, because her new knowledge does not concern facts. So, the no-fact reply provides a (common) way for materialists to protect materialism from the knowledge argument.

1.2.2 The Old-Fact Reply

As with the ability reply, all versions of the "old-fact" reply deny that Mary gains new factual knowledge. However, instead of denying that Mary's new knowledge is factual, they deny that Mary's new knowledge is about a *new* fact. As we will see, it seems that some facts can be represented in more than one way. The old-fact reply utilises this in order to claim that Mary gains a new way of knowing a fact that she already knew prior to her release. There are three main variations on the old-fact theme, which differ with respect to how they view Mary's new way of knowing.

The simplest version of the old-fact reply holds that Mary gains a new way of knowing (representing[6]) facts that she already knew (Churchland, P.M. 1995; Horgan 1984; Pettit 2004). Lots of things can be represented in more than one way. For example Eucalyptus trees can be represented in many different ways. We could take a photograph, describe the unique characteristics of the trees (such as their rigid, elongate leaves and flowers with many stamens and no petals), construct a model and so forth. However, the fact that there are

[5] Note that it is controversial to construe knowledge by acquaintance as non-factual knowledge. This is the basis of the most common criticism of Conee (Alter 1998). However my aim here is to neither endorse nor criticise these replies. Rather my aim is to survey the replies in order to get a sense of the lay of the land.

[6] As we will see below, whatever else is required for knowledge of a fact, representation of the fact is necessary (although not sufficient).

many different ways of representing eucalypts does not mean that there are many different things in the world. Rather, the one thing can be represented in more than one way.

Consider for example Paul Churchland's (1995) reply to the knowledge argument. He argues that phenomenal knowledge is "first-person" knowledge of one's brain. He draws an analogy between first-person knowledge of experience and first-person knowledge of one's bodily states. No one else can know about your body in the way that you know about it, using your proprioceptive system. This is simply because no one else is hooked up to your body in the right way. You have a privileged mode of access to your bodily states. However the existence of this special, first-person, way of knowing about your body does not imply that there is some different thing that is known. The fact that I can know about the positioning of my limbs via proprioception whereas you can't does not mean that what I know via proprioception is different from what you know via vision. In both cases we know about my body, we are just obtaining that knowledge in a different way. Similarly, it seems plausible that

> each one of us has a peculiar access, to exactly one's own sensations, that no other creature has. This is because each one of us enjoys a unique set of intimate *causal connections* to the sensory activity of one's own brain and nervous system. (Churchland, P.M. 1995, 196).

The idea is that, just as I have a special way of knowing about my body, so too do I have a special way of knowing about my experiences. However this does not mean that what I know via my internal pathways is something different from what the neuroscientist knows. The fact that Mary gains a new "introspective" way of knowing red sensations does not mean that the object of this "first-person" knowledge is different from the object of her "third-person" neuroscientific knowledge. There is simply more than one way of knowing the brain states that are colour experiences. The general idea behind the first type of old-fact reply is that Mary gain a new way of knowing facts that she knew prior to her release.

Some who claim that Mary's new knowledge is not about a *new* fact go further than simply claiming that she gains a new way of knowing facts that she already knew. They also claim that there are concepts about colour experiences that can only be possessed by someone who has had the experience (Loar 2004; Papineau 1993; Tye 2004). This is an intuitively plausible claim. It seems that the congenitally deaf person cannot even conceive of what sounds are like. Similarly, it seems that not only does Mary not know red; she cannot form a *conception* of red, or any other colour for that matter. What is important for our current purposes is that this "new concept" reply is still a version of the old-fact reply. The key to this reply is the claim that these new concepts are about the same, physical things that Mary knew prior to her release (Nida-Rümelin 2002). She gains a new way of conceptualising facts that she knew prior to her release.

The final variation of the old-fact reply holds that Mary's new knowledge is knowledge by acquaintance (Bigelow and Pargetter 1990; Churchland, P.M. 1985; 1989). Unlike Conee's no-fact acquaintance reply, this old-fact acquaintance reply holds that knowledge by acquaintance is a form of factual knowledge. For example Bigelow and Pargetter suggest that

> in attaining a new mode of acquaintance, Mary thereby opens the doors to a wide range of new thoughts which she can formulate ... Worlds which she could represent in one way, she can now represent in a new way as well (Bigelow and Pargetter 1990, 189).

The suggestion is that knowledge by acquaintance is a form of knowledge of facts. When Mary becomes acquainted with phenomenal colours, she gains a new way of knowing facts. She is now able to represent facts concerning phenomenal colours in a cognitively direct way. But, this new factual knowledge by acquaintance is not about new facts. Rather it provides her with new, direct, ways of knowing facts that she knew prior to her release.

Despite their differences, all of the versions of the old-fact reply share the idea that Mary gains a new way of knowing facts that she

knew in her black and white world. The central claim is that Mary does not gain knowledge of a new fact; rather she gains a new way of knowing an old fact. These replies secure materialism because if Mary's new knowledge is not about a new fact then it doesn't follow from the thought experiment that there are facts that are left out of her pre-release knowledge.

1.2.3 The Status of the Argument

The first thing to appreciate about both of these types of reply is that they protect materialism from the knowledge argument. The intuition that Mary will gain new knowledge is, like many pre-theoretical intuitions, imprecise. The standard replies make clear that the intuition that Mary will gain new knowledge is ambiguous and the argument only presents a problem for materialism on certain readings (Churchland, P.M. 1985; Horgan 1984 were first to point this out). There are, as we have just seen, many different ways of understanding the new knowledge that Mary gains in such a way that it does imply that Mary gains knowledge of a new fact. Importantly, we can now see that it is only if Mary gains knowledge of a new fact that materialism may be under threat. If Mary does not learn a new fact, then her pre-release knowledge is factually complete. So, there is one sense in which these standard replies are adequate. They protect materialism from the knowledge argument.

So, why are philosophers still so interested in Mary (as is demonstrated by the recent anthology of papers on Mary: Ludlow et al. 2004)? Why does the debate remain unresolved? I suggest that problem is that we do not know what would happen to Mary. For all their plausibility, we don't know whether any of the above replies are correct; or whether Mary will actually learn a new fact. This ignorance occurs simply because we do not know enough about the epistemology of conscious experiences to be able to demonstrate beyond reasonable doubt that one of the above types of reply is correct. If one of the above replies were true, then Mary would not present a problem

for materialism. Despite their plausibility, we do not know whether any of the accounts is correct. Given our uncertainty as to what would happen to Mary, the most decisive materialist response to the knowledge argument would be to cover all possibilities; to show that no matter what happens to Mary, materialism can still be true. However, as the standard replies do not attempt to deal with the possibility that Mary may learn a new fact, they are insufficient.

The situation is even worse than this. Not only do materialists tend not explore the possibility that Mary might learn a new fact, their tendency to deny that Mary learns a new fact actually reinforces the impression that *if* Mary learns a new fact then materialism is false. Even Jackson has capitulated and now thinks that Mary would not learn a new fact. Importantly, his reason for thinking this is that if Mary did learn a new fact then materialism would be false, and he thinks that letting philosophy revise science in such a way is a methodological mistake (Stoljar and Nagasawa 2004, 24). Despite Jackson's position, it is far from clear that materialism would be false if Mary learnt a new fact. Certainly we have the *intuition* that if Mary learnt a new fact then materialism would be false. Must this intuition turn out to be true?

The problem with Mary is not simply that she might be ignorant. Rather, the problem is that Mary is ignorant *despite* her (supposedly) complete physical knowledge. The (putative) fact that complete physical knowledge leaves out some knowledge suggests that these two sets of knowledge are about different things. If it turns out that Mary can know all the physical facts and not thereby know what colours are like, then the most obvious conclusion is that phenomenal colours are distinct from other physical things. She does not know about what it is like because in looking at physical things she is looking at the wrong things.

If we were sure that Mary knew all of the physical facts prior to her release, then there would be no reason to explore the possibility that she could acquire knowledge of new physical facts upon release. If Mary really did know all of the physical facts prior to release, and if she really did learn new facts upon release, then dualism would be

true. However Jackson explicitly avoids defining what it means for a fact to be physical (Jackson 1982, 127). This opens up the possibility that, despite our intuitions to the contrary, Mary may not know all of the physical facts prior to her release. Given the attention that has been paid to the claim that Mary gains knowledge of a new fact, it is somewhat surprising that so little attention has been paid to the claim that pre-release Mary has complete physical knowledge.

The primary purpose of this thesis is to see if facts that can be known only via experience can be physical. The motivation for examining the consequences of accepting that there are facts that can be known only via experience is to put the assumption that if Mary learns a new fact then materialism is false to the test. If the assumption turns out to be true, then at least we will know that materialists are right to focus their defence against the knowledge argument on denying that Mary learns a new fact. However, it may just be the case that materialism can still be true even if Mary learns a new fact. In the next section I will examine the replies of those few people who have denied that Mary knew all of the physical facts in order to home in on the problem with Mary.

1.3 What Is the Problem with Mary?

Although the majority of materialists have responded to the knowledge argument by denying that Mary learns a new fact, there are some who have attempted to demonstrate that Mary did not know all of the physical facts prior to her release. By examining their replies, and in particular by seeing how they are incomplete, we will be able to home in on precisely what remains to be done in order to demonstrate that materialism can be true even if Mary learns new facts upon release.

1.3.1 The General Strategy

Alter (1998) and Mandik (2001) have both presented accounts of the knowledge argument which attempt to show that materialism can still be true even if Mary learns a new fact. I will suggest that although their general approaches are correct, particularly concerning the epistemological side of Mary's situation, their accounts are too brief, particularly on important metaphysical issues, to be compelling.

Alter suggests that what Mary has before her release is not all of the physical facts, but rather all of the "discursively learnable" facts (1998, 50). According to Alter, what Mary has is all of the facts that can be learned by reading books and watching lectures on black and white TV. He then goes on to claim that there is no reason to think that the materialist is committed to the claim that all physical facts are discursively learnable. The materialist is not committed to any particular epistemic thesis about how things can be known. So, there is no reason for the materialist to accept the epistemic restriction that all physical facts be discursively learnable. Alter's claim is that the knowledge argument only refutes materialism if we take materialism to be not only the metaphysical thesis that everything is physical, but also the epistemic thesis that all physical facts can be learnt through discursive means.

This general point seems right: the materialist is certainly not committed to any epistemic restriction regarding the physical facts, such as that they are discursively learnable. Moreover it is certainly the case that if the knowledge intuition is true, then you cannot learn what an experience is like by reading books. The problem with Alter's account is simply its brevity. First of all, he does not go into any detail nor even provide an example of a non-discursively learnable fact. This is particularly worrisome because many pieces of non-discursively learnable knowledge are actually skills. For example you cannot learn how to play the piano or ride a bike simply by reading a book. Without providing a clear example of non-discursively learnable *factual*

knowledge, Alter does not clearly distinguish his reply from the no-fact reply.

The bigger problem with Alter's account, however, is the lack of a detailed metaphysical account of the physical yet non-discursively learnable facts. Materialism is a metaphysical thesis, so the materialist is not necessarily committed to any epistemological thesis concerning the nature of the physical. It nevertheless seems to be the case that, to this point in time, all physical facts have been discursively learnable. All of the obviously physical facts about the world, such as those dealt with by the sciences, can be learnt via discursive means. Although Alter is right to point out that there is no necessary connection between this epistemic profile and being physical, the onus is on the person who wants to break this association to provide a positive metaphysical account which demonstrates that there can be physical facts which can not be learnt via the usual discursive means. Without providing a more detailed metaphysical account of the facts about which Mary would learn, Alter has simply alerted us to the general way in which materialism could be true even if Mary learns a new fact. So, despite taking some first important steps towards suggesting that materialism can be true even if Mary learns a new fact, ultimately Alter's account is too brief to be convincing.

Pete Mandik's (2001) reply to the knowledge argument is more detailed that Alter's. But like Alter, Mandik fails to provide a sufficiently detailed metaphysical account of the facts that we can know only via experience. Nevertheless Mandik's basic epistemic account is correct. He suggests that we can understand Mary's inability to know some facts as an inability to represent those facts. One reason why there would be facts that we can *know* only via colour experience would be if there were facts that can be *represented* only via colour experience. This seems to more precisely capture Mary's problem than Alter's characterisation because it more clearly distinguishes Mary's new knowledge from skills.

In fact, understanding Mary's ignorance in representational terms seems quite obvious. Although epistemologists have yet to settle on an

account of factual knowledge, one thing that is thought to be necessary for knowledge of a fact is a representation of the fact (Steup 2006).[7] Given that representation is necessary for factual knowledge, one reason why we might be able to know a fact in only one way would be because we can represent it in only one way. Of course, more may be required to make this new representation new knowledge, but it seems clear that an inability to represent a fact will result in an inability to know the fact. Thus, I will follow Mandik and understand Mary's ignorance representationally. Mary learns new facts when she has colour experiences because there are some facts that she can only represent via colour experiences. Call these facts that Mary can know only by having colour experiences the "colour facts".

It is important to appreciate the restricted nature of the claim about these colour facts. The claim is *not* that the *only* possible way for the colour facts to be known is by having colour experiences. Rather the claim is that the only way for *Mary* to know the colour facts is by having colour experiences. The knowledge argument concerns what a normal human being, with all of their epistemic limitations, who has been prevented from having colour experiences, can know. In other words, the task is not to show that for all the possible knowers, colour experiences are the only way for them to know the colour facts. Rather, the task is to show that for a normal human being, colour experiences are the only way to know the colour facts. In other words, what we need to consider is whether the putative facts that a normal human being can know only by having experience can be physical. Note that for the sake of simplicity I will talk about the facts that *we* can know only by having experience. It is important to note that by "we" I mean normal adult human beings.

Mandik does more than simply suggest that we understand Mary's ignorance in representational terms. He also develops a representational account of her ignorance. In particular, he attempts to develop an account of facts, which can be represented in only one

[7] Although of course no one claims that representation of a fact is sufficient for knowledge of that fact.

way, which are nevertheless physical. Before we can understand Mandik's position we need to gain a better understanding of some basics regarding representation.

1.3.2 Representation

Representation can be found everywhere. The words on this page, the photographs on my wall, and the tank empty light on my dashboard are all representations. What these things have in common is that there is a *representing vehicle* that is *about* a *represented object.*[8] For example the word "tree" is about "perennial plants having a self-supporting woody main stem or trunk (which usually develops woody branches at some distance from the ground), and growing to a considerable height and size" (Oxford-English-Dictionary 2006). The photo on my wall is about the Sydney harbour bridge and the light on my dashboard is about the emptiness of my petrol tank. The representing vehicles are the string of letters "tree", the ink on photographic paper, and the light on my dashboard. The represented objects are perennial plants, the bridge in Sydney and the fact that the level of petrol in my tank is below a certain point, respectively.

One of the major tasks for any theory of representation is to account for how a representing vehicle can be about its represented object. This relation between the vehicles and object is sometimes known as the "grounding relation" (O'Brien and Opie 2004; Von Eckardt 1993). Theories of content determination (TCDs) provide accounts of how the vehicle gets to be about the object. Such theories appeal to three main types of grounding relations: causation, resemblance and convention. Although there are many different types of TCD, all theories invoke one or more of these types of grounding relations. The photo, for example, is about the bridge at least in part because it *resembles* the bridge. The photo preserves some of the spatial aspects

[8] Note that the use of the word 'object' to refer to what the vehicle is about should not be taken to mean that we can only represent concrete objects. The represented object can be anything; a concrete object, a property or an abstract object.

of the bridge: a picture taken side-on will, like the bridge when viewed from side on, depict two arcs being transected by a horizontal line just above the bottom of the upper arc with a series of vertical lines running from the outside arc through the inside arc to the horizontal line. In contrast the word “tree” is about trees because English speakers agree to use the word to refer to trees. It is a matter of *convention* that the string of letters refers to ‘a perennial plant having a self-supporting woody main stem or trunk' (Oxford-English-Dictionary 2006). Different linguistic communities have different conventions. For example French speakers use “arbre” to refer to perennial plants with a woody trunk. Finally the petrol light is about the quantity of petrol in my tank because it is *caused* to come on when the tank is empty. The light only comes on when the level of petrol in the tank no longer covers the sensor.

So far, we have determined that representation involves a representing vehicle that is about a represented object and that there are three general types of grounding relations. However there is more to representation than just a representing vehicle that bears a causal, resemblance or conventional relation to something. For example there are many more causal and resemblance relations in the world than there are representations. Everything (presumably) has a cause, and just about anything can resemble anything else (for example because they exist or because they are made of atoms). But, not everything is a representation. For example, drinking too much can cause a hangover, but a hangover does not seem to be a representation of excess drinking. Similarly, I resemble my sister Natalie, yet I do not seem to be a representation of her. So, what is it about the light on the dashboard and the photograph that distinguishes them from the hangover or me?

There doesn't seem to be anything special about the causal or resemblance relations themselves. Instead, what distinguishes the light from the hangover and the photograph from me is that we *make use* of the causal or resemblance relation between the vehicle and the object to give us information about the represented object. We use the light to indicate that the tank is empty and the photograph to, say show our overseas relatives the beauty of Sydney harbour. In fact, it seems that

we could make use of the apparently non-representational causal and resemblance relations in order to represent something. For example if I was trying to buy my sister a hat I could make use of the resemblance relation between myself and my sister to use myself to represent her. Our faces are of a similar shape, so if the hat suits me, it should also suit Natalie. Similarly, I could make use of the causal relation between drinking too much and a hangover to use the hangover as a representation of the quantity of alcohol consumed. For example I could infer from the severity of my hangover that the $50 that I spent last night went on drinks. The crucial point is that it is not enough for the vehicle to bear a content grounding relation to the object. What is also required is that the vehicle be used to stand in for the object. [9]

This making use of the relation between the vehicle and the object in order to use the vehicle to stand in for the object is called "interpretation" (Von Eckardt 1993). For non-mental representation, interpretation is 'the cognitive effect in the subject for whom the vehicle operates as a representation, such that this subject is brought into some appropriate relationship to the vehicle's object' (O'Brien and Opie 2004, 4). The role of interpretation is to mediate between the subject and the object such that the object of the representation can make a difference to the internal states and behaviour of the subject (Von Eckardt 1993, 157). For example, to be able to use the light on my dashboard as a representation of the emptiness of the petrol tank, I need to understand that it is caused by the emptiness of the tank. If I have no understanding of the relation between the light and the petrol tank I may wonder what the light is and puzzle as to why it only comes on just before I run out of petrol. But, if I understand that, when it is functioning properly, the light is caused to come on only when the tank is almost empty, then I can use the light to give me information about the state of my petrol tank in a way that changes my behaviour. When I see the light come on, I will make my way to a petrol station and fill up the tank. Understanding the relation between the representing vehicle

[9] This "triadic" account of representation is taken from Von Eckardt (1993)

(the light) and the represented object (the quantity of petrol in my tank) leads to a change in my behaviour that is appropriate to the object (filling up my petrol tank).

It is the existence of an interpretation that distinguishes the representational causal or resemblance relations from the non-representational causal or resemblance relations. It is only when I make use of the resemblance relation between my sister and I that I become a representation of her. Similarly, it is only when I make use of the causal relation between my hangover and the quantity of alcohol consumed that my hangover becomes a representation of the amount I drank the previous evening. On a triadic account of representation, (non-mental) representation involves not only a vehicle that is related by a content grounding relation to an object, but also a cognitive subject who is able to understand the grounding relation between the vehicle and the object in order to use the vehicle to stand in for the object. This is a genuinely triadic account of representation; each element (the vehicle, object and interpretation) is necessary and together the three elements are sufficient. There is no representation until the vehicle has been interpreted.

Although the above account works for non-mental representation, it will not work for mental representation. In particular, the interpretant (that is, the thing that does the interpreting) of mental representation cannot be the same as that for non-mental representation. For non-mental representation, interpretation is the *understanding* of a *subject*. For example, it is only when I *realise* that I resemble Natalie that I am able to use my face as a representation of hers. The problem with applying this account to mental representation is that this act of understanding seems to implicate representations. For example, it is by *thinking* to myself that I look like Natalie that I am able to use my face as a representation of hers. The problem with using mental representation as the interpretant of mental representation is that it leads to a regress (Von Eckardt 1993). As the interpretant is itself a representation, the interpretant will need to be interpreted. But, this second interpretant will itself be a mental representation and so will

require an interpretant that must be a representation, and so on. Given that the mind is finite, this sequence of representations/interpretants cannot go on forever. But, if it does stop somewhere, then the last representing vehicle will not have an interpretant and so it will not be a representation. But, if it is not a representation then it cannot be the interpretant for the representation preceding it, and so on until we find that the original vehicle is not interpreted and so is not a representation. It seems that the (ultimate) interpretant of a mental representation cannot itself be a representation because this would make the interpretation of mental representations impossible and therefore would mean, on the triadic account of representation, that there were no mental representations. As non-mental representations require mental representations as their interpretant, this would mean that there were no representations whatsoever. This looms as a reductio ad absurdum of a triadic account of representation.

As Von Eckardt (1993) points out, what we need is an interpretant of mental representations which is not itself a representation. Thankfully, there is an obvious candidate. It is a basic assumption of cognitive science that the purpose of having representations is to guide behaviour. For mental representation, interpretation can be making use of the relation between the vehicle and the object in order to guide behaviour. ‘A mental representation is interpreted in virtue of producing a change in the subject's mental life which, in turn, changes his or her disposition to act' (Von Eckardt 1993, 293). I do not need to represent the fact that the vehicle is related to the object in order to make use of the relation. Instead, all that is required is that I use the vehicle (perhaps indirectly) to guide my behaviour towards the object.

Note that not just any change in behavioural dispositions is sufficient. Rather

> the modifications must be in the person's tendencies to act in ways *dependent on the content of the representation*, for only such modifications will make a representation of some object O not only a representation *for* the subject but also a

> representation *of* O for the subject (Von Eckardt, 1993, p.284).

For example, reading “the cat sat on the mat” causes an increase in blood-flow to the areas of my brain responsible for the processing of written text. But, this increase in blood flow has nothing to do with the *content* of the phrase “the cat sat on the mat”. Reading the phrase “the boy chased the ball” could also cause it. Such an increase in blood flow does not count as the interpretant of “the cat sat on the mat”. This is because it is not dependent on the content of the representation. In contrast, getting up to vacuum the cat hair off the mat after reading the phrase “the cat sat on the mat” would count as the interpretant of the mental representation of the phrase. This is because the behaviour depends specifically on the content of the representation.[10] Reading “the boy chased the ball” would not prompt me to vacuum cat hair.

In general, the interpretant of a mental representation is the change in behavioural dispositions towards the object that the tokening of the vehicle enables. If tokening the vehicle leads to a change in my behavioural tendencies towards the object in a manner that is dependent on the content of the vehicle, then the vehicle has been interpreted.[11]

It is also important to note that not every interpretant must be a change in behavioural tendencies. All that is required is that that the regress is ended by an interpretant that is a change in behavioural tendencies (Von Eckardt 1993, 301). From any particular mental representing vehicle there must be some path to a behavioural modification. This means that the proximate interpretant of some mental representing vehicles can be a representation, as long as there is ultimately a path from the vehicle to a change in behavioural dispositions.

---

[10] Of course it is not necessary for one to actually start vacuuming in order for the representation to be interpreted. A change in behavioural dispositions is sufficient.

[11] Note that this account of interpretation does not make error impossible. There is no requirement that the change in behaviour towards the object be *appropriate*, just that it is dependent on the content of the vehicle. Error occurs when the vehicle is not about the object and so the change in behaviour towards the object is not appropriate.

So, on a triadic account representation involves three things - a *representing vehicle* that bears a content-grounding relation to a *represented object* and a representing system that *interprets* the vehicle by making use of the relation between the vehicle and the object to (ultimately) guide its behaviour towards the object.

1.3.3 Mandik's Account of Mary's Ignorance

Now that we have an understanding of the basics concerning representation we can turn to examine Mandik's account of Mary's ignorance. Mandik adopts an informational causal theory of content determination (akin to that developed by Dretske 1981; 1986) whereby a representing vehicle is about an object if it has the function to causally co vary with the object (Mandik 2001, 190). The basic idea behind an informational causal account of content is that, in veridical representation, the representing vehicle is about the represented object because it has the function of detecting the presence of the object in virtue of being causally related to the object. On such an account, my sweet experience of the chocolate I am eating is about the presence of sucrose in the chocolate because it is caused by the presence of sucrose in the chocolate and because those experiences have the function of detecting the presence of sucrose – it is their job to be caused by the ingestion of sucrose.[12] On such a theory of content, when I have the same sweet experience in response to an artificial sweetener, this is an instance of misrepresentation, because those experiences were not designed to detect the presence of those chemicals. The key point for now is that on such a causal informational theory of content, saying that there is only one way to represent something equates to saying that there is only one type of representing vehicle that has the function of being causally related to the object. If there is only one 'detection supporting causal interaction' (Mandik

[12] This is an example of a TCD that utilises causation as a grounding relation.

2001, 198) that the fact can enter into, then there will be only one way of representing that fact.[13]

In order to explain how there can be physical things that have only a small number of detection supporting causal interactions, Mandik uses Dennett's (1991) Jell-O box example. In order to have a safe way of recognising each other, spies would tear something, such as a Jell-O box, in half. Tearing the box produces an edge of such complexity that it is virtually impossible to reproduce. The edge of one half becomes a unique pattern-recognition device for the other half. Mandik suggests that 'the only *detection supporting* causal interactions that one piece of the Jell-O box enters into are causal interactions involving the other piece of the box' (2001, 198).

It seems possible that some of the properties represented in sensation are complex properties that only enter into representation supporting causal interactions with certain types of sensory equipment. This would mean that someone without that sensory equipment would be unable to represent those properties. The idea is that I would not be able to represent those features that a bat represents in its echolocatory experiences because, lacking their echolocatory sensory equipment, I simply cannot have the requisite representations caused in me. I lack the sensory equipment required to have the representations that a bat has. Therefore I cannot represent the properties that can only be represented in those representations.

Applying this to Mary, the claim is that there is a physical property that she could not represent until she had her first colour experience 'because the property enters into detection supporting causal interactions only with the mechanisms of human colour vision' (Mandik 2001, 199). The idea is that the mechanism that produces colour experiences is the only mechanism that can causally interact with the

[13] Van Gulick also suggests that there may be "subjective" facts that can be represented in only one way because of causal limitations in the type of representations any individual can have. 'Once the informational or representational content of a theoretical structure is seen to depend on the way in which a subject's causal nature enables him to use that theoretical structure, the suggestion that the resources of physical theory are adequate for representing all facts becomes implausible' (Van Gulick 1985, 67)

facts that are represented in colour experience and therefore is the only mechanism that can produce representations that are about those facts.

I want to suggest that Mandik's commitment to a causal-informational theory of content determination is not necessary. The general point is that the reason why there would be facts that can be represented only via colour experience is because colour experiences are the only mental representations that can have the necessary content-grounding relation to the colour facts. On a TCD that utilised causation as a grounding relation, this would be because the only mental representing vehicles that the colour facts can cause are colour experiences. But, on a TCD that utilised resemblance as a grounding relation, we would be ignorant because colour experiences are the only mental representing vehicles that can *resemble* the colour facts. In general, we would be ignorant because there is only one type of mental representing vehicle that can have the appropriate content-grounding relation to the colour facts. Mandik's key idea is that there are facts that can be represented in only one way because there is only one type of representing vehicle that can be about that fact. This reply can be made regardless of what you think it is that makes a representing vehicle be about a fact.

So far, what Mandik has shown is that the reason why there would be facts that we can know only via experience would be because there are facts that we can represent in only one way. The reason why there are facts that can be represented in only one way is because there is only one type of representing vehicle that can bear the appropriate content grounding relation to the colour facts. The harder task, on which Mandik trips up, is to provide a detailed metaphysical account of these facts that we can only represent in colour experience which clearly demonstrates that they are physical.

Demonstrating that facts that we can know only via experience are physical is no small task. As we have already seen, it seems that obviously physical things can be known in many different ways. It doesn't seem necessary to have a particular type of experience to

learn, say, quantum physics. This seems to suggest that any facts that we can know only via experience cannot be physical.

Mandik begins his account of the facts that we can only represent in colour experience by arguing that they "depend for their existence" on the representations that can only be about them (Mandik 2001, 197). He makes this point via an analogy with pictorial representations. Consider for example the photo in figure 1.1, which I took whilst visiting to Japan. Mandik suggests that pictures, such as this one, include within them information about the spatial location of the viewer (in this case the camera). He also suggests that the facts that are picked out in such representations depend for their existence on these pictorial representations. Consider for example all of the surfaces in the garden that were reflecting light to the lens of the camera. Mandik suggests that what groups these surfaces together into a coherent whole is the fact that they are all and only those surfaces which were reflecting light onto the camera lens. The surfaces exist independent of the representation of them, but it only makes sense to group them together relative to the point of view of the camera. They only cohere as a group relative to the representational point of view from which they can be represented as a group. Take away the point of view and there would be no reason to group them together. What unifies all the visible points into a set is the point of view of the viewer: 'What the image represents depends for its existence on the process of representation in the sense that precisely what is represented cannot be characterised independently of specifying the point of view of the representing subject …[because] the point of view occupied by the viewer … is the unifying essence of those particular regions' (Mandik 2001, 197). In other words, according to Mandik, we cannot pick out the relevant surfaces without specifying the point of view of the observer. The elements depend for their existence as a coherent whole on the representation that represents them as a whole.

Figure 1.1: Pictorial perspective in a Japanese garden.

There is undeniably something correct about this. In particular, it is clear that what unifies the surfaces that are represented in the picture in figure 1.1 is that they are the ones which were reflecting light to the lens of the camera. But, and here is the problem, it is far from clear that this collection of surfaces is some *new* unified fact that is distinct from the set of surfaces. Although it is clear that what unifies the surface is the image, it is not clear that the "unity" that is represented is real. Why is the set of surfaces anything more than just a collection? Why is it a *new* fact that could not be represented by just providing a (very long) list of the relevant bits of foliage and water? Mandik argues that:

> Of course there is a sense in which they [the surfaces] would exist even if no one were to represent them. But in specifying the set comprised of all and only the spatial regions captured in the image, one does not carve nature at the joints, but instead carves nature into a gerrymandered collection of items that would be of no interest apart from

> their involvement in a particular representation (Mandik 2001, 197)

Although it seems reasonable to say that we may not be carving nature at its joints in pictorial representation this is in fact the problem. If we are not carving nature at its joints, then in what sense are the things that we are representing in these images, or colour experiences *in nature?* It is an open possibility that some of the ways we represent the world as being are not actually ways the world is. For example, it is possible that colours as we see them are not actually properties of objects in the world, but rather things that we project *onto* the world. What Mandik needs to provide is an account of these facts that are represented in colour experience that shows that the facts themselves are somehow more than the collection that makes them up and that our representations of them as distinct facts are not illusory. It is not enough to show that we represent them as more than their parts, because it is an open possibility that some of the ways in which we represent the world are not actually ways in which the world is.

Mandik devotes very little attention to demonstrating that the facts that are represented in experience really are distinct from the collection of surfaces that constitute them. This brevity is a main flaw with his account. He suggests that 'there is one sense, I will grant, that we can represent all of the physical properties from the objective point of view, and another sense in which we cannot.' (Mandik 2001, 199). In order to explicate the senses in which we can and cannot represent all of the physical facts from the objective point of view he uses the example of his French friend, Jacques Blanc. He suggests that there is a sense in which the linguistic representation “all of the people in France are less than 3 meters tall”, represents something about Jacques (who is of average height), and that there is a sense in which it doesn't.

> The sense in which it does is the sense in which it says something true of Jacques Blanc. The sense in which it does not is the sense in which it is not specifically about Jacques - it does not single out Jacques from the other inhabitants of France (Mandik 2001, 199).

Mandik suggests that something similar holds for sensory experience in that 'the properties delivered in sensory experience are quite specific' (2001, 200). Sensory experience supposedly provides us with a way of representing more specific facts than physical theory. However, Mandik argues that this in no way gives us any reason to think that such facts are not physical. The inability of physical theory to represent these facts is no more mysterious than the inability of physical theory to provide us with the resources to recognise one half of the Jell-O box without the other, or the fact that “all of the people in France are less than 3 meters tall” is not specifically about Jacques.

Although there is nothing necessarily wrong with this account, it seems to be inadequate. For example, as we will see in chapter two, we know quite a lot about the nature of the facts that are represented in colour experience. More importantly there is currently no obvious bar to learning all that science can tell us about these facts without actually having colour experiences. Without providing a more detailed account of the specific facts that are represented in sensory experience, Mandik has failed to meet his primary aim, namely showing that such facts are physical. The specific sensory facts are only marginally less mysterious than the facts that we can know only via experience. I do not mean to suggest that such facts cannot be physical. All I am suggesting is that without a more detailed positive account, the claim that they are physical and yet can only be represented in experience is not much more than a statement of faith. Mandik has failed to provide a detailed metaphysical account of the facts that we can know only via experience, so, he has failed to demonstrate that such facts can be physical. Although Mandik makes some crucial first moves, ultimately his account is not much more than a promissory note.

Although Mandik’s basic account of *why* there are facts that we can known only via experience is excellent, he has not provided enough detail of *what* we are ignorant for it to be clear that what we do not know is nevertheless physical. What we can see from Mandik’s reply to the knowledge argument is that the unresolved problem with Mary is that she may learn a new fact by having colour experiences. In

particular, the problem with the Mary thought experiment is that it can be interpreted as suggesting that there are facts that Mary can know only via colour experience. Although Mandik has provided an account of why Mary does not know, what is lacking is an account of *what* she doesn't know that is clearly consistent with materialism.

As we have seen, there are numerous other interpretations of what happens to Mary, all of which are compatible with materialism. However, there is one interpretation of what happens to Mary, namely that she learns a new facts when she has colour experiences, which has not yet been clearly demonstrated to be compatible with materialism. The task of this thesis is to see if facts that can be represented only via experience can be physical.

You may have noticed that so far, not much has been said about consciousness. In the next section I will consider what we can say about consciousness given that we have accepted that there are facts that we can know only via experience.

1.4 What About Consciousness?

The knowledge argument is concerned with showing that consciousness cannot be physical. But, so far we have had to say very little about consciousness. In fact, if the replies to the knowledge argument are anything to go on, then the argument seems to concern knowledge much more than consciousness. In this section I want to briefly consider what we can say about consciousness, given that we have accepted that there are facts that we can represent only by having colour experiences. In particular I will focus on two issues: 1) what can we say about the relation between consciousness and representation given that we have accepted that there are facts that can be known only via experience; and 2) what do we need to say about consciousness given that the intuition is that the only way to know what an experience is like is to have the experience. We will see

that we can in fact put what it is like to one side whilst dealing with the knowledge argument.

It should be obvious that if Mary is to learn new facts, then colour experiences need to be representations. The reason why there are facts that Mary can't represent is because colour experiences provide the only way to represent those facts. This is consistent with the general consensus: it is generally accepted that most experiences have representational content. Although some argue that pain or orgasm are not representational, everyone is happy to admit that visual experiences are representations (Block 1990; 1996; Lycan 2005). As the focus of the knowledge argument is colour experience, this is crucial. It is uncontroversial to say that visual experiences are representations, and this is part of what we need in order for there to be facts that we can represent only via experience. The requirement that colour experiences be representations is entirely uncontroversial. However we also need it to be the case that colour experiences have a representational content that is distinct from other experiences.

If colour experiences are to be the *only* way for us to represent some facts, then not only do colour experiences need to be representations, they also need to have a different representational content than other experiences. If auditory experiences or achromatic visual experiences can have the same content as colour visual experiences, then there are no facts that we can represent *only* via *colour* experience. So, part of what is required for Mary to learn new facts is not only that experiences are representations, but also that different experiences have different representational content. What does accepting this tell us about consciousness?

One obvious way to ensure that different experiences have different representational content is if the phenomenal content of an experience is identical to its representational content. This is the key idea behind "representationalism". According to the representationalist, the phenomenal qualities of experience are held to be 'intentional contents, represented properties of represented objects.' (Lycan 2005,

3). The idea is that the what it is like of an experience comes from the way in which we are representing the object of experience.

Representationalism comes in weak and strong forms (Lycan 2005). According to the strong representationalist, all phenomenal content is to be identified with representational content (Dretske 1995; Tye 1995; 2000). Weak representationalists deny that phenomenal content can be exhaustively understood in terms of representational content (Block 1990; 1996; Shoemaker 1994).

The key point for the current discussion is that weak representationalism is sufficient for there to be facts that can be represented only via experience. As long as some aspect of phenomenal qualities are representational, then to change the phenomenal quality is to change its representational content. This association between representation and consciousness is accepted by weak representationalists. For example, Shoemaker thinks that weak representationalism is consistent with the claim that a difference in phenomenal character leads to a difference in representational content. He writes: 'within the experiences of a single person, sameness or difference of qualitative character will go with sameness or difference of representational contents' (Shoemaker 1994, 22). Similarly, Block, a paradigmatic weak representationalist, thinks that 'it is often the phenomenal character itself that has the representational content' (Block 1996, 20). If the phenomenal character "has" the representational content, then presumably different phenomenal characters will "have" different content. Importantly, colour experiences are *not* one of the types of experiences for which the weak representationalist claims that the representationalist identification of phenomenal quality and representational content fails (Block 1990; 1996; Lycan 2005).

So, representationalism, in either its weak or strong form, ensures that different experiences have different representational contents. The important thing about the compatibility of Mary's ignorance with weak representationalism is the almost universal acceptance of representationalism (Lycan 2005, 4). When it comes to consciousness,

accepting that there are facts that we can represent only via experience has simply told us what most already believe.

If Mary is to learn a new fact, then representationalism needs to be true. But, what sort of facts does Mary learn about? It is standard to hold that what Mary learns about is what colour experiences are like. The facts that Mary can only represent by having colour experiences seem to be the phenomenal facts. If this is correct, then to demonstrate that Mary learns new physical facts by having colour experiences, one would need to demonstrate that phenomenal colours are physical. This is asking a lot. There is a school of thought that holds that there is a "hard problem" of consciousness. The hard problem is that there is something uniquely difficult about demonstrating that phenomenal experiences are physical (Chalmers 1996; McGinn 1989; Nagel, T. 1974; 1986; 1998; Searle 1992). In other words, to show that Mary's new knowledge of what colour experiences are like is knowledge of physical facts, one would need to solve the hard problem of consciousness. This is somewhat beyond the scope of this work.

Instead, the tactics that I have adopted from Mandik (2001), are more indirect. The knowledge argument is supposed to demonstrate that phenomenal qualities cannot be physical because they can be known only via experience. To answer this challenge the materialist does not need to demonstrate that phenomenal colours are the physical facts that can be known only via experience. Rather they need to demonstrate that facts that we can know only via experience can be physical. The aim is to do this by developing an account of colour vision on which there are physical facts that we can represent only by having colour experiences. Whether these facts are phenomenal facts is a question for another day. If I succeed, I will not have shown that what an experience is like is physical, even though the only way to know it is via experience. Rather, I will have shown that the (putative) fact that the only way to know what an experience is like is to have the experience is not evidence that what it is like is non-physical, because there can be physical facts that can be known only via experience.

This is appropriate given the supposed relation between the knowledge argument and the hard problem of consciousness. The knowledge argument is supposed to provide evidence in support of the existence of a hard problem. So, dealing with the knowledge argument should not require us to solve the hard problem of consciousness. It is only by putting what it is like to one side that we can get at the central issue of the knowledge argument, namely determining the metaphysical status of facts that we can know only via experience.

The aim of this thesis is to develop an account of colour vision on which there are physical facts that we can know only via colour experiences. If this can be achieved, then the knowledge argument will be disarmed. Even if it turns out that the only way to know what an experience is like is to have the experience, this will not longer be evidence that what it is like is not physical. This is because it will have been shown that there are other *physical* facts that we can know only via experience. If I succeed there will no longer be an epistemic difference between what it is like and other physical facts.

1.5 Overview

We can now see that the Mary thought experiment only secures dualism if: (a) Mary gains knowledge of a new fact and; (b) facts that we can know only via experience cannot be physical. If either of these claims is false, then Mary presents no problem for the materialist. The point of this thesis is to provide a plausible account on which (b) is false; on which the facts that we can know only via experience are physical. The aim is to show that materialism can be true *even if* Mary learns a new fact, such that the dualist can no longer simply pump the intuition that Mary would be ignorant in order to support their position.

We have seen that Mandik, and to a lesser extent Alter, have made some important first moves in accounting for Mary's ignorance. In particular, Mandik's suggestion that the reason that there are facts that we can know only via is because colour experiences are the only

representing vehicles that can bear the appropriate content grounding relation to these facts, also seems to be correct. What we still need is a detailed metaphysical account of the facts that we can only represent via colour experience. One which shows that such facts are physical. The task of this thesis is to provide such an account.

The first thing we need to do is to consider the nature of the facts that are represented in colour experience: the “colour facts”. This is the task of the second chapter. We will see that the psychophysical data concerning colour presents a great challenge to the claim that colour experiences are representational. Basically, colour appearance varies with so many factors that it is far from clear that colour experiences are representations. However, I will argue that the dependence of colour appearance on the spectral, spatial and figural features of the entire visual field is not mysterious. Rather it is the means by which the visual system overcomes the ambiguity of the retinal image to achieve a more accurate representation of the reflectance of a surface. I will argue that if we adopt a molecular scheme of representation it is possible to provide a representational account of colour experience that does justice to the psychophysical data. We will see that colour experiences are part of our visual model of the world and that they represent (at least) the spectral, spatial and figural properties of the world.

It is one thing to show that what we are representing in colour experience is complex phenomenon concerning the relation between the spectral, spatial and figural features of the world. It is quite another to demonstrate that these complex states of affairs are metaphysically distinct from the spectral, spatial and figural features that constitute them. This is precisely what is required if there are to be facts that we can know *only* via colour experience. What we need to show is that the complex states of affairs that are represented in colour vision are metaphysically distinct from their constituents. That the colour facts are mereologically emergent. In order to do this it will be necessary to develop a general, scientifically motivated metaphysics. In the third chapter I will argue that the metaphysics that correctly captures what science is telling us about the world is an emergentist one.

I will begin by surveying the various notions of emergence that have been developed and will suggest that the two core notions are that emergent phenomena are mereologically novel and inexplicable in mereological terms. I will then turn to consider some scientific examples to see whether we can find either mereological novelty or inexplicability. We will see that although science is full of examples of mereologically novel phenomena, these phenomena are explicable in mereological terms. I will develop a metaphysics on which there is explicable, emergent, novelty. We will see that the key is to base the novelty of a complex phenomenon in the physical organisation of its parts. Then I will argue that this "structural" account of emergence avoids the problems with causation that have plagued the popular version of non-reductive materialism that bases irreducibility of phenomenon in multiple realisability. Finally, I will argue that the states of affairs that are represented in colour experience are "structurally" emergent and therefore are metaphysically distinct from their constituents.

Once I have developed the proper metaphysical account of the colour facts I will turn, in the final chapter, to consider how this constrains the epistemology of these facts. In particular, I will examine whether it is possible to develop an account of visual versus non-visual representation on which we can only represent the colour facts visually. We will see that it is plausible that it is only in vision that we have the representational resources to represent the colour facts in their entirety. Although we may be able to non-visually represent various aspects of the colour facts, I will suggest that only vision has the capacity to represent the emergent phenomena that are the colour facts.

The task of accounting for the facts that we can only represent in colour experience will then be complete. The facts that are represented in colour experience are complex, emergent states of affairs concerning the relation between the spectral, spatial and figural properties of the world. We can only represent these facts via colour experience because only the visual system has the resources to

represent such complex facts. Although it is possible to represent the constituents of these states of affairs non-visually, I will suggest that we do not have the resources to construct a representation of any colour fact in its entirety except in colour vision. Because these colour facts are structurally emergent, the whole is metaphysically distinct from the parts. Therefore, non-visual knowledge of the constituents of the colour facts is insufficient for knowledge of the colour facts as a whole. So, there are facts that we can know only via colour experience.

I will finish by examining the status of materialism in light of this account of colour vision. Can facts that we can know in only one way be physical? Does accepting that there are facts that we can know only via experience mean that we can't know that materialism is true? I will argue that, as with other structurally emergent phenomena, the colour facts are physical. I will also argue that an inability to know the facts themselves does not prevent one from knowing that they are physical, nor does it mean that our science is necessarily incomplete.

The outcome will be a demonstration that facts that we can know only via experience can be physical. This means that, regardless of what happens to Mary, materialism can still be true. We will see that the real importance of the knowledge argument is not that it has the potential to prove that materialism is false, but rather that accounting for its intuition forces us to push the boundaries of our understanding of the mind and the world.

Chapter Two
The Represented Object
Of Colour Experience

In this chapter I will begin to develop a model of the facts that we can represent only via colour experience.[14] In particular, the task of this chapter is to develop an account of the facts that are represented in colour experience. Before we can try to account for why there are facts that we can represent *only* via colour experience, we need to develop an account of the *facts* that are represented in colour experience. This is no small task. Colour vision has been studied intensively by scientists and philosophers. Despite a wealth of data we still have no clear idea what colour experiences represent. In fact, the appearance of a surface varies with so many factors that it has been claimed that colour experiences do not represent at all (Mollon 1989; Thompson 1995; 2000).

The most intuitively obvious account of colour experiences is that they represent the spectral features of objects (Byrne and Hilbert 2003a; 2003b). On such an account our phenomenal blue experiences represent surfaces which reflect photons with a wavelength of around

[14] I want to thank Jon Opie for his advice on colour. I first came across the idea that colour experiences might represent more than surface spectral reflectances in his cognitive science course. My subsequent discussions with him on colour have been invaluable.

450nm and our phenomenally red experiences represent surfaces which reflect photons with a wavelength of around 700nm. However the psychophysical data does not fit easily into this account. Various psychophysical results demonstrate that the appearance of a surface depends on the spectral, spatial and figural nature of the entire visual scene. A number of theorists have proposed that this dependence is an aspect of the visual system's constancy mechanism. However this does not in and of itself tell us what, if anything, is represented in colour experience. Do colour experiences represent the relations between the spectral and spatial aspects of the world? Or do they represent surface spectral reflectances, partly by using spectral, spatial and figural information? Ultimately the answer to this question will be informed by one's theory of representational content.

In this chapter I will argue that adopting a molecular scheme of representation provides a new way of thinking about the content of colour experiences. In particular it enables us to develop a representational account of colour vision that can accommodate the complexities of the psychophysical data.

In the first section I will examine the psychophysical data that seems to be inconsistent with the claim that colour experiences are representations. The second section will focus on why it may be the case that the spectral, spatial and figural properties of the world are represented together. In the third section I will introduce the distinction between molecular and lexical schemes of representation and show how adopting a molecular scheme of representation provides a new way of thinking about the represented object of colour experiences. We will see that colour experiences are part of our visual model of the world that represents (at least) the spectral, spatial and figural properties of the world.

Before moving on to the main task of this chapter, we need to clarify a few issues. First of all we need to gain a clearer understanding of what a fact is, then we need to draw some distinctions between the types of fact concerning colours.

As with everything in philosophy, there is no consensus regarding what facts are. Particularly amongst more analytically inclined philosophers, it is common to understand a fact as a true proposition (Frege 1984; Lowe 1995; Sosa 1995). I want to suggest that given that our goal is to develop a *scientifically* motivated metaphysics, and given how complicated propositions are and how difficult it is to think of them metaphysically,[15] it is perhaps best to avoid construing facts as true propositions. Instead, I suggest that it is best to understand a fact as an actual state of affairs (Russell 1918; Sosa 1995). Of course, this immediately raises the question of what a state of affairs is. This is another question whose answer would require a thesis of its own. However I think we can get by with the rather vague understanding of a fact (or state of affairs) as simply a way the world is (Alter 1998).

The deeper problem for the current discussion is that "colour fact" is ambiguous. In particular we can distinguish between three different types of "colour facts". First of all, there are the "clearly objective"[16] states of affairs involving surfaces and light sources, which our visual system responds to, that are investigated by physicists. As we will see below, a colour fact in this sense might be that the part of the wall in front of me that is behind the filing cabinet is painted with the same paint as the rest of the wall; it is just reflecting light with a lower intensity because the filing cabinet is casting a shadow on it. Secondly there are facts concerning phenomenal colour experiences. Facts about the organisation of our phenomenal colour space – such as the fact that red and green are opponent hues, or that colour space is not symmetrical because the most saturated yellow is lighter than the most saturated blue (Palmer 1999a) – would be colour facts in this

[15] For example it is an open question as to how to individuate propositions. Some people think that two propositions are the same if they are true in all the same possible worlds (King 2006). On such an account of propositions, all necessary truths (such as 2+2=4 and bachelors are unmarried) express the same proposition because they are all true in all possible worlds.

[16] 'Objective' is meant to indicate external to the mind/brain. 'Clearly' is meant to leave open the possibility that there are some facts concerning colours beyond those investigated by sciences such as physics which might also turn out to be external to the mind. In particular, the intent is to avoid begging the question against the phenomenal externalist.

phenomenal sense. Finally, there are facts concerning the relation between the clearly objective colour facts, and the phenomenal colour facts. Call these the "relational colour facts." This is an area that has attracted a lot of philosophical interest.[17] A colour fact in this sense may be that phenomenal redness is identical to a particular surface property or that surfaces with a particular reflectance when illuminated with a particular type of light will cause phenomenally red experience.

So, what sorts of colour facts are we interested in when investigating the represented object of colour experiences? Obviously we are interested in the clearly objective facts about colour, the things in the world which colour experiences seem to be about.[18] But, simply examining the clearly objective facts about the world will not help us determine what it is that human beings with normal colour vision are representing in their colour experiences. In order to answer this question, the key point to appreciate is that, as I have argued in the previous chapter, if there are to be facts that we can represent only via experience, then we are committed to representationalism about colour experiences. Recall that if there are to be facts that we can represent only via experience then the phenomenal content of colour experiences must be their representational content. This means that in order to determine what we are representing in colour experiences we need to examine the relation between our phenomenal colour experiences and the clearly objective features of the world. That is, given that we have accepted representationalism, an examination of the relational colour facts holds the key to determining what it is that we are representing in colour experiences. So, the methodology of this chapter will be to focus on the psychophysical data concerning the relation between our colour experiences and the clearly objective features of the world.

[17] See for example Akins (1996b), Akins and Hahn (2000) Arstila (2003), Averill (1992), Byrne and Hilbert (2003a; 2003b), Churchland (2005), Cohen (2003), De Anna (2002), Dedrick (1996), Hardin (1988; 1990; 2003), Hilbert (1992), Hilbert and Kalderon (2000), Kraut (1992), Matthen (1999), Maund (2006), McGinn (1993), Revonsuo (2001), Rosenthal (2001), Ross (2001a; 2001b), Sinnott-Armstrong and Sparrow (2002), Thomas (2001), Thompson (1992), Turvey et al (2001) and Webster (2002).

[18] Although of course it is in fact the burden of this chapter to argue that colour experiences are representations of things in the external world.

Secondly, a few notes to situate myself in the rather large philosophical literature concerning the metaphysics of colour. In philosophy the hot topic is the debate between the colour objectivists and subjectivists. According to colour objectivists, colours are a property of the mind-independent world. According to colour subjectivists, colours are a property of the mind. Given the distinctions between the different types of colour facts drawn above, it should be obvious that subjectivism and objectivism as defined above are not unambiguous positions. In particular, it is important to distinguish between phenomenal and non-phenomenal objectivism/subjectivism.[19] On the one hand there is the question of whether phenomenal colours, colours as they appear to us, are in the head, or in the world. On the other hand there is the question of whether colours simpliciter are in the head or in the world. Perhaps the easiest way to appreciate that there is a distinction to be drawn is to consider an analogy with shape (Akins and Hahn 2000). Objects rarely appear to be the shape that they actually are. For example the rectangular top of my desk currently appears to be a trapezium and the circular top of the cup on my desk appears to be an oval. It is natural to draw a distinction between how shapes appear and how they really are. Moreover, once we draw this distinction we can see that the metaphysical questions concerning phenomenal and objective shapes can be given different answers. For example we can claim that phenomenal shapes are in the head whereas objective shapes are properties of things in the world. Similarly, once we draw a distinction between phenomenal colours and objective colours we open up the possibility of answering the metaphysical questions about them in different ways.

[19] What I am terming "phenomenal colour subjectivism" is usually known as "phenomenal internalism" and what I call "phenomenal colour objectivism" is usually known as "phenomenal externalism".

Note that it is, arguably, misleading to identify the phenomenal internalism/externalism distinction with one way of drawing the subjectivism/objectivism distinction. The explicit focus of the objectivism debate is not phenomenal colours (Maund 2006). However, as Akins and Hahn's (2000) analogy with space makes clear, the phenomenal and objectivism questions are quite distinct. Given how much of the debate over colour objectivism focuses on phenomenal colours, I think it is important to be careful to keep questions concerning the metaphysics of phenomenal colours versus colours simpliciter distinct.

The reason why I am drawing this distinction between phenomenal and non-phenomenal colour objectivism is that the psychophysical data I use below strongly suggests that phenomenal colours are not properties of the objective world. The problem is that there seems to be no mapping from colour experiences to an objective property. This suggests that colour experiences do not correspond to some objective property, and therefore that *phenomenal* colours cannot be identified with an objective property. The psychophysical data seem to suggest that phenomenal subjectivism (internalism) is correct (Hardin 1988; 1990; 2003 has long argued for this).

The important point is that if phenomenal subjectivism is true, and *phenomenal* colours are a property of the mind, this does not mean that colours simpliciter must also be mind-dependent. As Akins and Hahn (Akins 1996b; Akins and Hahn 2000) have argued, the absence of a mapping from phenomenal experiences to an objective property is not evidence for (non-phenomenal) subjectivism. They point out that

> While size and shape are surely objective properties in the world, radio towers *look* small … when viewed from a distance and tables almost never produce squarish experiences except on those unusual occasions when we occupy a viewpoint directly above them (or, more commonly, under them). Moreover, when a one-to-one correspondence between a phenomenal appearance and its distal cause fails to hold, we do not treat such cases as threatening the putative objectivity of those properties … So, the co-variation of sensory data or phenomenal experiences with a mind-independent distal cause is not a necessary condition of objectivity (Akins and Hahn 2000, 233-4).

Just as the absence of a mapping from our experience of size and shape to an objective property does not mean that size and shape are not objective features of the world, so too does the absence of a mapping from our experience of “colours” to an objective property not mean that “colours” are not objective. Although I am committed to

phenomenal subjectivism (internalism), this does not commit me to subjectivism for colours simpliciter. Just as we need to draw a distinction between the metaphysics of phenomenal and objective shapes, it is still an open possibility that there are objective features of the world that deserve to be called colours. The upshot is that I can avoid entering the quagmire that is the debate between colour objectivists and subjectivist. The current task is to investigate the relation between the phenomenal colour facts and the clearly objective features of the world in order to determine what feature of the world colour experiences represent.

2.1 The Psychophysical Data

The common-sense account of colour experience holds that colour experiences are about the material properties of objects. When I say that the ripe tomato looks red, it seems that I am making a claim about the nature of the surface of the tomato. The most obvious property that these claims would be about is surface spectral reflectances (SSRs).[20] In high school I was taught that the tomato looks red because it reflects photons of a long wavelength and absorbs photons of a shorter wavelength. The fundamental reason why it has proven so difficult to provide an account of the facts that are represented in colour experiences is that the appearance of a surface depends on more than the spectral nature of the surface. Colour science has shown us that common sense and my high school science teacher are wrong. Colour experiences vary with the spectral, spatial and figural nature of the entire visual field. The challenge is to provide an account of the represented object of colour experience that is consistent with this psychophysical data.

[20] The surface spectral reflectance is the percentage of incident light that a surface reflects at each wavelength.

2.1.1 The Spectral Nature of the Surround

It has long been known that the spectral nature of its surround plays a crucial role in determining the appearance of a surface. Perhaps the most persuasive evidence in support of this is the fact that black[21] and brown are what is known as contrast colours. What this means is that black and brown can only be experienced when the surface that appears black or brown is surrounded by a surface of contrasting colour. A surface that appears brown when surrounded by contrasting regions will look orange or yellow when it fills the visual field. Similarly, a surface that appears black when surrounded by contrasting regions will look grey or even luminous when it fills the visual field (Gelb 1938; Kaiser and Boynton 1996, 46; Pokorny et al. 1991, 46). The fact that a surface can go from appearing luminous to appearing black, or from appearing yellow to appearing brown, simply by changing the nature of its surround, suggests that what we are representing in colour experiences is something more than the light being reflected from the surface itself.

The spectral nature of the surround plays an important role in determining colour appearance for all colours, not only black and brown. In simultaneous colour contrast, the appearance of a surface is influenced by the spectral nature of its surround. In particular, two surfaces with identical SSRs can appear to be different colours because they are surrounded by regions with different reflectances. For example, in figure 2.1, the two central squares have been printed with the same ink. However they appear to be different shades of blue because of their different surrounds. If they are being illuminated with a constant source of light, the spectral distribution of the light reaching the retina from both of the central squares should be identical. Yet they appear to be different colours. This suggests that colour appearance

[21] Although the achromatic shades are not always thought to be “colours” they are part of our colour space and so a discussion of them is appropriate in any discussion of colour vision.

depends not only on the nature of the light leaving the surface, but also on the spectral nature of its surround.

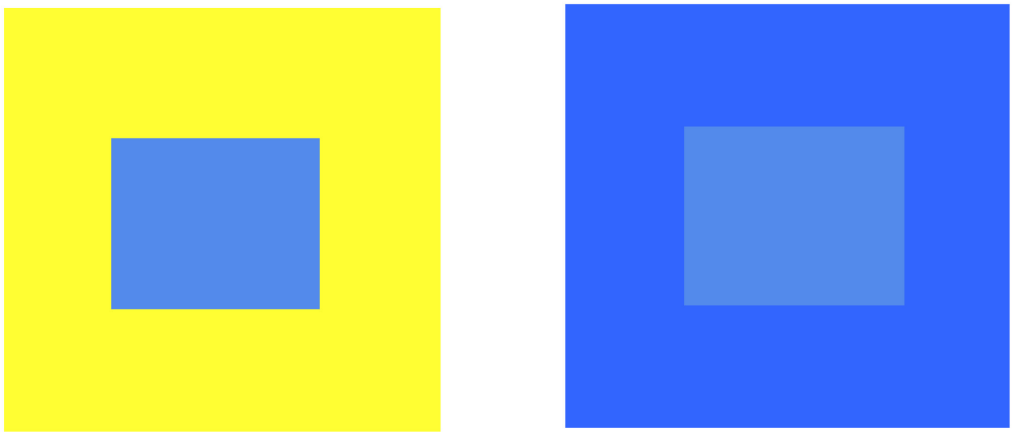


Figure 2.1: Simultaneous Colour Contrast. The two centre squares have the same reflectance but they appear to be different shades of blue because of spectral differences in their surround.

What can contrast colours and simultaneous colour contrast tell us about the represented object of colour experience? Because simultaneous colour contrast involves representing two identical surfaces as different it is tempting to write it off as an illusion; as an aberration that can be accommodated once we have the correct theory of colour vision. We are representing two identical surfaces as different, so it seems that we must be making a mistake. This temptation is misleading. First of all, although the central blue regions in figure 2.1 are the same, their surrounds are different. Although the surfaces are the same, they are part of two *different* complex figures. It is not the case that the *stimuli* themselves are identical. Rather, two different stimuli have identical components that appear to be different. In fact, simultaneous colour contrast is only an illusion on the assumption that colour experiences represent surface spectral

reflectances. If colour experiences are about SSRs, then any case where surfaces with the same SSR are represented as different is an instance of misrepresentation. However, whether or not colour experiences represent SSRs is precisely what is up for grabs. To call simultaneous colour contrast an illusion, without argument, is to beg the question.

More importantly, the existence of contrast colours provides a reason for thinking that contrast is central to colour vision. As Hardin points out

> To write off simultaneous contrast as something that need not enter into one's fundamental theory of colours is also to write off the possibility of giving a proper account of the nature of black and brown. This seems unacceptable, unless one is prepared to think of black and brown as "illusory" rather than "real" colours (1990, 560).

The fact that the presence of a surrounding region is necessary in order to have an experience of black or brown suggests that the colour experiences are representing more than the reflectance of an individual region. As Hardin points out, unless every experience of black and brown is to be an illusion, colour experiences must represent something more than the reflectance of a surface.

In response to Hardin, and in defence of objectivism, Byrne and Hilbert (2003a; 2003b) suggest that we need to distinguish between the conditions necessary for the perception of a property and the property that is perceived. They argue that all that can be concluded from the existence of contrast colours is that the presence of a contrasting surround is a necessary condition for the perception of such colours. They draw an analogy with the effect that the nature of the surround can have on spatial properties. Consider, for example, the Müller-Lyer illusion (see figure 2.2). In this illusion, two lines of equal length appear to have different lengths because of differences in their surround. The analogy with contrast colours is obvious. In both instances objects with identical properties are represented as distinct because of differences

in their surround. Byrne and Hilbert suggest that we need to treat colour and space equally:

> Illusions such as the Müller-Lyer illusion do not show that the length of an object is relative to its surround. Rather they show that sometimes an object's surround can prevent us from seeing the length it actually has …. Similarly, the conditions necessary for the perception of colour include facts about the surround, and also facts about mechanisms within the brain of the viewer. But this does not show that colour is relative to the surround, or the perceiver, or that no external object has a colour (Byrne and Hilbert 2003b, 39).

Although this distinction seems appropriate and may be useful in defending objectivism, simply drawing the distinction leaves unanswered how the visual system makes use of such information. More particularly, it leaves unanswered precisely what is being *represented* in any particular colour experience.

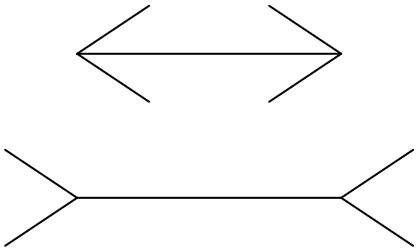


Figure 2.2: The Müller-Lyer Illusion. The two horizontal lines are the same length. However, they appear to be different lengths because of spatial differences in their surrounds.

Changing the spectral nature of its surround can change how a surface is experienced. This suggests that the visual system is taking into account the spectral nature of its surround when determining how to represent the surface. Similarly, the Müller-Lyer illusion suggests that the visual system takes into account the nature of the surround when representing length. Although the property the system is *trying* to

represent in colour experience may be SSR, it may be the case that in order to obtain a more accurate representation of SSR the system needs to take into account more factors than just the material properties of an individual surface. Similarly, although the property that the visual system is trying to represent in spatial visual experiences is the objective spatial properties of the figure, it may be the case that in order to obtain a more accurate representation of the objective spatial properties it is necessary to take into account the spatial properties of the rest of the visual field. As we have already learned from Akins and Hahn (2000), issues concerning representation and those concerning the metaphysics of colour are independent. I will have a lot more to say on the mechanisms for achieving constancy below. For now I just want to point out that even if Byrne and Hilbert's distinction is valid, simply drawing the distinction leaves unanswered how it is the case that the "conditions necessary for perception" are taken into account. In particular, they do not rule out the possibility that the conditions are part of the representational content of the experience.

It seems that the visual system takes into account the spectral nature of a region's surround when determining how to represent the surface. However the spectral nature of the surround is not the only factor that is important in determining colour appearance.

2.1.2 Space, Figure and Colour

The appearance of a surface is affected not only by the spectral properties of its surround. The spatial and figural properties of the surfaces are also important. The most pure example of the importance of spatial properties in determining colour appearance is the effect of relative size on brightness. In a series of experiments Li and Gilchrist (1999) explored the effect of the relative size of two achromatic regions on their apparent brightness. They put subjects in a Ganzfield[22] and

[22] A Ganzfield is visual field produced by set-up where a (usually quite simple) stimulus fills the entire visual field. In this experiment subjects were placed underneath a large acrylic hemisphere.

split the field up into two achromatic regions of different luminosity (either by dividing the circle into two sections or by presenting a disc surrounded by an annulus). They found that the perceived lightness of the stimuli varied with the relative size of the regions, even though only the spatial and *not* the spectral properties of the stimuli were changed. In particular it was found that as the size of the less luminous (darker) region increased, both regions appeared lighter until the smaller region appeared luminous and the darker region appeared almost white. Simply by increasing the size of a surface relative to its surround the surface can go from appearing black to a very light grey.

The spatial complexity and structure of a stimulus also plays an important role. In fact, by increasing the spatial complexity of the stimuli an effect that is opposite to that of the standard colour contrast effect can be found (see figure 2.3). Colour assimilation occurs in stimuli that are more complex and iterated than the standard colour contrast stimuli. What is interesting about assimilation is that the surround has the opposite effect to that which it has in simultaneous colour contrast. In simultaneous colour contrast (figure 2.1), the blue square that is surrounded by the blue region appears *less* blue than an identical square that is surrounded by yellow. In colour assimilation (figure 2.3), the blue squares that are surrounded by blue appear to be *more* blue than those that are surrounded by yellow. In general, in colour contrast the appearance of the target is shifted *away* from that of its surround, whereas in colour assimilation the appearance of the target is shifted *towards* that of its surround. Simply by increasing the spatial complexity of the stimuli, the effect of the surround is reversed. This finding suggests that the effect of the surround is mediated by the spatial complexity of the stimulus.

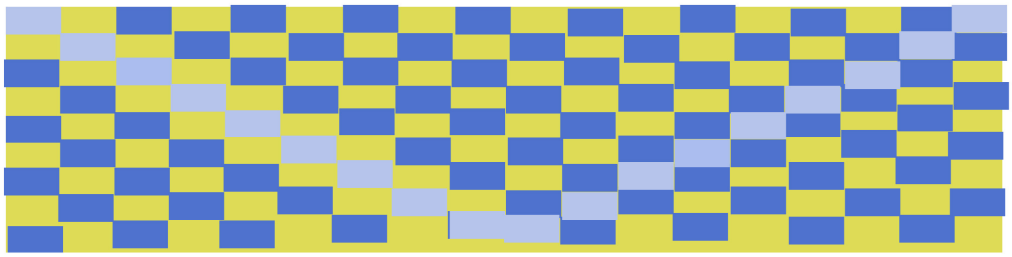


Figure 2.3: Colour Assimilation. The grey/blue squares along the diagonal are printed with the same ink, but they appear to be difference colours. In particular, their appearance is shifted towards that of their immediate surround. So, the squares on the left that are surrounded by yellow squares appear yellower (less blue) whereas those on the right that are surrounded by blue squares appear bluer.

The perceived three-dimensional layout of surfaces in the visual field also affects colour appearance. In both colour contrast and colour assimilation it is the nature of the immediate surround that affects the appearance of a surface. For example the target squares on the right of figure 2.3 look bluer because of the influence of their immediate blue surround. They do not appear to be greatly influenced by the more distant yellow squares. Although the nature of the immediate surround is important, there is evidence that more global comparisons, particularly those which are based on perceptual belongingness, are also crucial (Agostini and Galmonte 2002; Gilchrist et al. 1999; Li and Gilchrist 1999).

For example, Agostini and Galmonte (2002) have demonstrated that a surface is contrasted with the surfaces that it is perceived as belonging with, not necessarily its immediate surround. They constructed Necker cubes (figure 2.4) where the local contrast and

figural contrast cues were opposing. We can see that the middle lines of each side of the cube consist of dashed lines of an intermediate grey. The corners of the cube are black or white and the background is black or grey such that the shade of the background is opposite to that of the corners.

Figure 2.4: Figural Cues and Contrast. These Necker Cube Stimuli demonstrate the role of figural factors in the perception of achromatic colours. The dashed grey lines are printed with the same ink; however they appear to be different shades because they are contrasted with their surrounds. Importantly, they are contrasted not with the background, which is their immediate surround, but rather with the corners of the cube, that is, with the regions they are perceived as belonging with (after Agostini and Galmonte 2002, 90)

The logic behind this experiment is that if contrast only occurs locally, then the central dashed lines should be contrasted with the background. However, if contrast is determined by global factors, such as the object that the surface is perceived as belonging to, then the central dashed lines should be contrasted with the corners of the

"cube" to which they appear to belong. Because the background and the corners are opposite shades, they will have opposing contrast effects on the central lines.

Agostini and Galmonte found that the central lines were contrasted with the corners, not the background. Recall that contrast has the effect of moving the appearance of the target away from the appearance of its surround. So, the grey that is contrasted with a lighter region will appear darker than that which is contrasted with a darker region. We can see that the central lines in the figure on the left appear to be lighter than those on the right (even though they are actually the same shade of grey). This suggests that the lines on the left are being contrasted not with the lighter background, but rather with the darker corners. In other words, Agostini and Galmonte found that the target grey lines were contrasted with the regions they appear to belong with, rather than their immediate surround. This suggests that for complex scenes, the figural properties of the stimulus determine which regions are contrasted with each other, and therefore in part determine the appearance of the region.[23]

In general, it seems that the appearance of a surface depends not only on the spectral nature of its surround but also on the spatial and figural properties of the visual field as a whole. Identical surfaces can appear to be different because of spectral, spatial or figural differences in their surround. These findings create a number of problems for the claim that colour experiences are representational.

2.1.3 The Challenge for a Representational Account of Colour Experience

Data such as that presented above have been used to argue that colour experiences are not representations (Mollon 1989; Thompson 1995; 2000). There are two main problems. The first difficulty that the

[23] For a further discussion of the effects of scene complexity ('articulation') on contrast see: Gilchrist (1977), Gilchrist et al (1999), Kraft et al (2002), Maloney and Schrillo (2002) and Neumeyer et al (2002).

psychophysical data raises is that there does not appear to be any simple mapping from a particular colour experience, say a red experience, to some objective property of surfaces – “redness”. If you manipulate the lighting conditions then just about any surface can appear red. But, if just about anything can appear red, then it is not clear that red experiences are actually representing anything. The second problem arises from the fact that we are considering what colour *experiences* represent. Science seems to be telling us that colour experiences represent more than the properties of a particular surface. But, it appears that our experiences are about individual surfaces. How can we be so wrong about the content of our *experiences*? Aren’t experiences just how things appear? If colour experiences appear to be about surfaces, then how can we say that they are not?

The fact that, with appropriate manipulations of the lighting conditions, just about any surface can appear red suggests that red experiences are not representations. The absence of a mapping from experiences to an objective property, presents a challenge to the claim that colour experience are representations. If just about any surface can appear red, then how can we avoid saying that red experiences represent *every* surface? Moreover, given that just about any surface can appear blue (if the lighting conditions are appropriately manipulated), then how can we distinguish between what red experiences represent and what blue experiences represent? Of course, misrepresentation and illusions occur. Sometimes we use a representing vehicle to stand in for something it is not actually about. So, in searching for the represented object it would be a mistake to try and include *everything* that the experience is used to represent. Some of these will be instances of misrepresentation. Nevertheless, the veridical and non-veridical cases need to be separated. The problem is that it is hard to see how this can be done unless the representational question has already been answered.

Two main suggestions for separating veridical and illusory colour experiences have been made. First, Shepard (1992) has suggested

that we can define normal conditions in terms of natural light. The idea is that perception is veridical under normal conditions and illusory otherwise. For example a representation of a banana as yellow would be veridical because this is how it appears under normal daylight. Its orange appearance under a red light would be classified as illusory because red illumination is not normal. Unfortunately this is problematic. There is enormous variability in the wavelength of light across the day (Palmer 1999b). Importantly there is more than enough to bring about a noticeable difference in the appearance of a surface.[24] Worse still, there is also evidence that colour constancy is *worse* in naturally occurring lighting conditions than in artificial lighting conditions (Brainard 1998; Delahunt and Brainard 2004; Ruttiger et al. 2001). So, it is not clear that we can use a definition of normal conditions in terms of natural daylight to separate the veridical from the non-veridical colour experiences.

It has also been suggested that we can distinguish between veridical and non-veridical experiences on teleological grounds. In particular, the proposal is to focus on what the system has been designed to detect (Hilbert 1992; Matthen 1988; 1999; 2001; Mollon 1989). According to this account, veridical experiences would be those that represent what the system has been designed to represent. A major problem with such an approach is the inference from function to content. Although the system may have the *function* of detecting surface spectral reflectance, it may be the case that in order to do this it actually needs to represent something more than the reflectance of the surface.[25] In fact, in the next section we will see that it is plausible that in order to gain accurate information about SSRs it is necessary to take into account the spectral, spatial and figural properties of the entire visual scene.

[24] Although this usually does not lead the surfaces being put into a different colour category. The banana will appear to be different shades of say yellow across, but will usually not go from looking yellow to looking red.

[25] This inference may be valid on teleological theories of content (see for instance Dretske 1981; Millikan 1989).

It is far from clear that we can find a principled way to separate the veridical from illusory colour experiences. In the absence of such a separation, the fact that there is no mapping from red experiences to some objective property of surfaces suggests that colour experiences are not representational. Perhaps we can't separate out the illusory from the veridical instances because they are all illusions.

The other problem that faces representational accounts of colour experience is the tension between introspection and science. Introspection tells us that colour experiences are about individual surfaces. Psychophysics tells us colour experiences depend on more than the nature of the light leaving the surface. How can colour experiences *appear* to be about specific surfaces if the appearance of a surface depends on the entire visual scene? This is particularly puzzling when we realize that we are trying to answer the question "what do colour *experiences* represent?" How could we be so systematically wrong about how things appear to us in experience? Of course, I do not want to deny that we can make mistakes about the nature of our experiences. What I do want to suggest is that it would be nice if we could avoid attributing such a radical mistake to an entire form of experience. At least some colour experiences seem to be about the properties of surfaces. If we take the psychophysics at face value then the appearance of a surface depends on many things besides the nature of the surface. Could we really be that wrong?

I want to suggest that it is possible to develop an account of the represented object of colour experience which can make sense of the psychophysical data and do justice to introspection. To understand this account we first need to consider why it is the case that colour appearance is determined by the spectral, spatial and figural nature of the surround.

2.2 A Complex Visual Model

2.2.1 Achieving Constancy

The fundamental problem of vision (Hoffman 1998) is that the retinal image is ambiguous. Many different distal stimuli could have caused the one retinal image. For example, the information that the visual system receives concerning the reflectance of a surface is fundamentally ambiguous. The nature of the light reflected from a surface (the luminance spectrum) depends not only on the reflectance of the surface but also on the nature of the illuminating light (Palmer 1999b, 122-4). Two surfaces with the same reflectance can reflect light with different spectra because they are illuminated with light with different spectra (think of moving a white piece of paper from a blue to a red spotlight). Moreover, two surfaces with different reflectances can reflect light with the same spectra if they are illuminated with the appropriate lights (think of looking at a red piece of paper under a white spotlight and a white piece of paper under a red spotlight). So, in order to represent surfaces as having the same reflectance despite changes in illumination, the visual system needs to "undo" the effects of illumination on the luminance spectrum. But, 'the problem is under-constrained because the eye receives information about only one proximal variable (luminance) which is determined jointly by two distal variables (reflectance and illumination)' (Palmer 1999b, 125). In principle we cannot determine whether it is a white piece of paper under a red spotlight, or a red piece of paper under a white spotlight. A particular luminance spectrum could have been caused by an infinite number of combinations of illuminants and surface reflectances.

Despite the ambiguity of the retinal information we are able to achieve a high level of illumination-independent constancy (Gilchrist et al. 1999, 796). We are able to represent a surface as the same despite changes in the light reflected from it due to a change in the illumination. For example the text on this page appears black whether it is viewed

indoors or in full sunlight, even though the spectrum of the reflected light may vary dramatically in the two conditions. In fact, if viewed under a particularly dim light, the white background of the page may reflect less light than the black text does when it is in direct sunlight (Palmer 1999a, p.125). This type of constancy is not perfect, as the white paper under the red spotlight example demonstrates. Nevertheless we are able to achieve a reasonable level of illumination-independent constancy. This means that the visual system must have mechanisms that enable it to (partially) overcome the inadequacy of the information it receives. By understanding the heuristics the visual system uses to deal with the ambiguity of the retinal information, we can begin to appreciate why the appearance of a surface depends on the spectral, spatial and figural nature of its surround.

Although we still lack a complete account of the means by which the visual system is able to achieve a constant representation of the reflectance of a surface, a comparison of the light being reflected from different regions is one mechanism that is almost universally appealed to (Gilchrist et al. 1999; Land, E.H. and McCann 1971; Lotto and Purves 2000; 2002; Rock 1975; Wallach 1948; Williams et al. 1998a; 1998b).[26] Such a comparison is useful because a luminance change will often affect all surfaces equally. Although the actual amount of light reflected by a surface varies depending on the nature of the illumination, the reflectance of the surface is constant. This means that under the same illumination a darker surface will reflect less light than a lighter surface. Although the white background of this page, when viewed in dim light, may reflect less light than the black text when

[26] Note that a comparison of the light being reflected from adjacent surfaces is not the only proposed means of achieving illumination-independent constancy. For example it has also been suggested that it can be achieved by the adaptation of receptors (Churchland, P.M. 2005; Hills and Brainard 2005; Maloney, L.T. 1986; Robilotto and Zaidi 2004; Smithson and Zaidi 2004) or by a direct assessment of the illumination (Brainard and Wandell 1986; Cunthasaksiri et al. 2004; Maloney, L. T. 2002; Smithson 2005). The debate on which of these mechanisms is used still rages (for instance Rutherford and Brainard 2002, claim to have falsified the illumination-estimation hypothesis). It is likely that the visual system uses as many cues as possible, and so it is likely that all of these sources of information are utilized (Gegenfurtner and Kiper 2003; Kraft and Brainard 1999). So it is important to realize that I am not in any way claiming that a comparison of the luminances of adjacent regions is the *only* way that illumination-independent constancy is achieved.

viewed in bright sunlight, the ratio between the reflectances of the page and the text will stay the same across changes in illumination (Rock 1975, 507). Under the same illumination the white background will always reflect more light than the black text. Similarly, a blue piece of paper will always reflect more short wavelength light than a red piece of paper. In general, when illumination changes, neighbouring regions will be affected in the same way such that comparing the reflectances of neighbouring regions can help the visual system to factor out the effects of the illuminant. This has led many theorists to suggest that comparing the reflectances of adjacent regions is part of the visual system's illumination-independent constancy mechanism.

What is important for our purposes is that if illumination-independent constancy is (at least partially) achieved via such a mechanism then the fact that the appearance of a surface depends on the spectral nature of its surround would be explicable. In particular, it is part of the means by which the visual system is able to overcome the poverty of the information it receives and achieve a constant representation of the reflectance of a surface.

However, as Gilchrist, Kossyfidis et al (1999) point out, contrast cannot be the only determinant of the appearance of a surface. If contrast were the only factor, then we would not be capable of background-independent constancy. Background-independent constancy refers to the fact that we are able to represent a surface as being the same despite changes in its background. We find the colour contrast and assimilation figures so surprising because we usually do not see a surface as changing colour as its background changes (you can test this by moving a coloured object around so that it is being contrasted with different surfaces). Although surrounding a surface with regions with different SSRs can change the appearance of a surface, this is not something that occurs in an extreme way outside of the laboratory. In fact, as I discovered when constructing the contrast and assimilation figures above, it is reasonably difficult to generate striking examples of these phenomena. Although contrast plays a role, it does not rule the roost. There needs to be other mechanisms which assist

the visual system to achieve not only illumination-independent but also background-independent constancy.

One of the things that the visual system needs to do in order to achieve both background and illumination independent constancy is to determine whether a change in the light being reflected from a surface is due to a change in illumination or in SSR. Drawing such a distinction is important because a correction for a change in illumination when the change is due to a change in surface reflectance would lead to the misperception of the reflectance of the surface.

There are various heuristics that the visual system can use in order to distinguish between reflectance and illumination changes. For example, illumination edges tend to be fuzzy and somewhat graded, whereas reflectance edges tend to be sharper (Palmer 1999b). You can see this for yourself by comparing a shadow to a painted edge. A comparison of the luminance difference between two regions is also useful. Illumination changes produce much greater changes in luminance than reflectance changes. For example a typical white surface reflects 90% of incident photons and a black surface only reflects 10%. So, the luminance difference for reflectance edges is unlikely to be more than about 10:1. In contrast the luminance difference for illumination ratios can be up to 1000:1 (Palmer 1999b, 132).

Spatial cues are also useful in distinguishing between reflectance and illumination changes. This is because they can provide information about the nature of the illumination striking a surface. For example, if a surface is perpendicular to a light source it will receive maximum illumination; the more a surface is slanted from the source, the less light it will receive. Similarly, the further away from a light source a surface is the less light it will receive (Rock 1975, 525). Surfaces at different depths and/or orientations also usually receive different amounts of illumination. So, if depth information indicates that two regions are not coplanar, an edge between them tends to be perceived as an illumination edge rather than a reflectance edge. Determining the relative positions of surfaces and the illuminant can assist the visual

system in detecting any differences between the illumination reaching two surfaces. Cues as to the three-dimensional lay-out of objects can also assist the visual system in determining where shadows are likely to be cast and therefore in determining where there are likely to be changes in illumination (Lotto and Purves 2002). Figural cues can also help. For example, if two surfaces are part of the same object they are likely to be similar distances from the light source and therefore are more likely receiving the same amount of light than if they belong to different objects (Agostini and Galmonte 2002).

This suggests that the dependence of the appearance of a surface on the spatial and figural properties of its surround is part of the means by which the visual system achieves both background-independent and illumination-independent constancy. In fact, we can understand the results of Agostini and Galmonte's (2002) study by considering the role of spatial cues in distinguishing illumination from reflectance changes. Recall that they found that a surface was contrasted with the surfaces it was perceived as belonging with, not with its immediate surround. We can now appreciate the logic behind this result. Surfaces of the same object tend to be receiving similar illumination, because their distance from the illuminant tends to be the same. The nature of the Necker cube illusion is such that the background is perceived to be at a different depth to the sides of the cube. As surfaces at similar depths are more likely to be receiving the same illumination, it makes more sense to contrast the dashed lines with the other "sides" of the cube. Because they are more likely to be under the same illumination there is less chance of the luminance difference between them being due to both a reflectance and illumination difference. In contrast, as they are perceived as being at different depths, it is quite likely that the luminance difference between the dashed lines and the background is due to both an illumination *and* reflectance difference. Because of this, it is likely that a comparison of the dashed lines with the background would confound the illumination and reflectance difference between them. So, the dashed lines are not contrasted with the background. What we can see here is that the

visual system uses figural cues in order to determine which regions are likely to be in the same plane, and therefore more likely to be receiving the same illumination, in order to decide which regions should be contrasted. In other words, the visual system uses figural cues in order to try and distinguish between reflectance and illumination differences.

In general, spatial and figural information provides cues that can assist in determining the nature of the illumination striking a surface. This suggests that instead of being an indication that colour experiences are not representations, the dependence of the appearance of a surface on the spectral, spatial and figural nature of its surround is part of the visual system's means of retrieving as much information as possible from the ambiguous retinal information. The nature of the light reaching the eye is determined by (at least) the nature of the illumination, the material properties of the object and the spatial arrangement of the surfaces and illuminants. Given that the nature of the proximal stimulus is determined by the relations between these distal properties, it makes sense to model these relations in order to more accurately represent the distal properties. This suggests that spectral, spatial and figural properties are represented together in order to increase the accuracy of the representation of the world. This claim receives further support from the fact that taking spectral properties into account helps achieve a more accurate representation of the spatial properties of the stimulus.

The retinal information concerning spatial properties is also ambiguous. The retinal image is two-dimensional whereas the world is three-dimensional. So, there is an inevitable loss of information and ambiguity as to the distal cause of the retinal image. Importantly, there is some evidence that suggests that spectral information can assist the visual system to achieve some measure of spatial constancy. Spectral cues are implicated in many models of shape and object perception (Fleming et al. 2004; Lehky and Sejnowski 1988; Palmer 1999b) and there is a growing body of evidence that object perception is more accurate for scenes in which there is chromatic variation (Edwards et

al. 2003; Gegenfurtner and Rieger 2000; Kingdom 2003; Rossion and Pourtois 2004; Tanaka et al. 2001).

One mechanism of shape perception is to determine shape from shading (Horn and Brooks 1989; Lehky and Sejnowski 1988; 1990; Marr 1982; Ramachandran 1988). The intensity of the light reflected from a surface varies depending on its orientation from the source of illumination. A surface that is perpendicular to a light source will be receiving more light than one that is parallel to the illuminant. This means that by comparing the intensity of the light reflected from different surfaces, it is possible to determine their relative orientations. If the position of the light source is known, it is also possible to determine the absolute positioning of the surfaces. In general, shading information can provide information as to the three-dimensional shape of objects.

Given our current purpose, the problem with determining shape from shading is that at first glance, shading information does not seem to require colour vision. What distinguishes colour vision from achromatic vision is the ability to distinguish between light of the same intensity but with different spectral distributions. However, shading information concerns differences in *intensity* not differences in the wavelength. However, this appearance is misleading. Wavelength information is useful in distinguishing between illumination and reflectance changes and therefore in identifying shadows. For example, a change in illumination will usually only cause a change in the intensity, but not the spectral distribution of the reflected light. By looking for a change in spectral distribution, that is, by considering wavelength information, the visual system can determine whether an edge is a reflectance or illumination edge and therefore decide whether to take the edge into account when calculating shape from shading. Importantly, there is experimental evidence that wavelength information assists in the perception of shadows (Kingdom 2003).

Not only do spatial and figural cues assist in achieving a more accurate representation of the spectral properties of the world, it also appears that spectral cues can assist in achieving a more accurate

representation of the spatial and figural properties of the world. What the example of shape from shading demonstrates is the co-dependence of the information that the visual system receives about different properties in the world. The nature of the light reaching the eyes is determined by the spatial layout of objects and light sources as well as the spectral nature of objects and light sources. Given the ambiguity of the retinal information, it makes sense to model the co-dependence of these features when representing the world. By allowing the representations of things that interact in the world to mutually constrain each other, the possible interpretations of the retinal image are reduced. So, the dependence of colour appearance on the spatial and figural properties of a surface is not necessarily evidence for the claim that colour experiences are not representational. Instead, it appears to be part of the means by which the visual system deals with the ambiguity of the retinal information and achieves a more accurate representation of the spectral and spatial properties of the world.

2.2.2 Additional Empirical Considerations

Further support for the claim that there is a co-dependence between the way in which spectral, spatial and figural properties are represented can be found in the fact that colour experiences capture the "statistical regularities" of visual scenes (Long and Purves 2003; Lotto and Purves 2002; Nundy and Purves 2002; Yang and Purves 2004). For example, Long and Purves (2003) analysed images of natural scenes in order to determine the effect of the surround on a particular region. The aim was to see if context effects, such as colour contrast and assimilation, could be explained in terms of the typical environmental cause of such stimulus configurations. They calculated the probability that a particular region would have a certain spectral configuration, given the spectral characteristics and spatial complexity of its surround. They found that for natural scenes with a more simple spatial structure, similar to the standard contrast stimuli, a surface in a green context tends to be

redder, whereas a surface in a red context tends to be greener. In other words, for simple, contrast-like stimuli, the spectral return of the target tended to be shifted *away* from that of its surround. In contrast, for natural images with a more complex spatial structure, similar to that of assimilation stimuli, they found that the spectral return of the target was shifted *towards* that of the background. So, for highly iterated stimuli, a surface on a green background tended to be greener whereas one on a red background tended to be redder.

The interesting thing about this data is the similarity between the statistical analysis of images and the psychophysical data considered earlier. Recall that in contrast stimuli, the appearance of the target is shifted away from the surround. Long and Purves found that in images of natural scenes with a spatial complexity similar to that of contrast stimuli, that the spectral return of the target tends to be shifted away from that of the surround. Similarly, in colour assimilation, the appearance of a region is shifted towards that of the surround. In their analysis of images, Long and Purves found that in scenes that were spatially similar to assimilation stimuli that the spectral return of the target was shifted towards that of its surround:

> In general, the perceived colours of the targets in colour contrast, constancy, and assimilation stimuli are all predicted by the way that their positions are shifted in subjective colour space by the typical co-occurrence in natural scenes of the spectral returns of the targets together with the relevant spectral context and spatial configuration (Long and Purves 2003, 15193).

In other words there is a similarity between the spectral *and spatial* structure of natural scenes and the relations between colour experiences such that the shifts in colour space that various types of surround induce in a target resembles the spectral returns of the target and its surround in natural scenes. This finding, that colour experiences resemble the spectral and spatial properties of natural scenes, supports the claim that the spectral and spatial properties of the world are modelled together.

One major empirical challenge for the suggestion that there is a co-dependence between the representations of spectral, spatial and figural information is the fact that an influential view of the visual cortex holds that there are separate visual pathways for the processing of information about different visible properties such as colour, depth, form and motion. In their attempt to link psychophysical and physiological studies, Livingstone and Hubel (1987; 1988) suggested that the visual system can be divided into two main pathways: the magno and parvo. According to this view, colour and form are processed together in the parvo system until V1, where the parvo system splits into separate colour and form systems. The magno system is supposedly specialized for depth and motion, which are also processed together until V1, after which they are processed separately.

The functional descriptions of the magno and parvo systems are based on microelectrode studies. In these single cell studies, a microelectrode is inserted into a cell and the receptive field of a neuron is mapped by determining the pattern of retinal activity that causes the highest rate of firing. For example, the magno system is said to be colour-blind, quick, with high contrast sensitivity and low resolution. This is because magnocellular layer cells in the Lateral Geniculate Nucleus (LGN) do not distinguish between equiluminant stimuli of different wavelengths; because they tend to have larger receptive fields than parvo LGN cells; because they respond faster and more transiently than parvo cells; and because they are much more sensitive to low-contrast stimuli (Livingstone and Hubel 1988, 741). Similarly the parvo system region V4 is usually reported as being the colour centre because in a series of microelectrode studies of the cortex Zeki found that cells in V4 respond selectively to particular wavelengths (Zeki 1993, 125).

Further support for the dissociation of various visual functions can also be found in psychophysics, where a dissociation between colour vision and depth and motion vision has been demonstrated. What distinguishes the colour system from the achromatic system is that it can distinguish between equiluminant stimuli of different wavelengths.

So, demonstrations that motion or form perception is not possible in stimuli where the only variation in light is wavelength ("equiluminant stimuli") are an indication that wavelength ("colour") information is not used in the processing of these visual attributes. If the visual system used wavelength information in the processing of motion, then we would be able to see movement in isoluminant stimuli. There is a considerable evidence that the perception of motion is diminished or absent at equiluminance (Carney et al. 1987; Cavanagh et al. 1984; Livingstone and Hubel 1987; 1988; Ramachandran 1987; Ramachandran and Gregory 1978), and that depth perception is degraded at equiluminance (Livingstone and Hubel 1987). This suggests that wavelength information is processed separately to motion and depth information.

The agreement of the physiological and psychophysical data presented above provides powerful evidence in favour of the claim that colour is processed separately from motion, depth and form. More importantly, it calls into challenge the current claim that the spectral, spatial and figural properties of the world are represented together. Fortunately there are a number of problems with this simple view of the visual cortex.

First of all the physiology is not as simple as Livingstone and Hubel's position requires. For example the notion that V4 is "the" colour centre is called into question by a number of recent studies (Hanazawa et al. 2000; Wachtler et al. 2003). In fact, in a recent review, Gegenfurtner and Kiper suggest that:

> although a consensual estimate is still lacking, it is likely that V4 contains approximately the same proportion of colour-selective cells as V1, V2, or V3. Moreover, the colour selectivity of V4 cells appears to be similar to that of neurons in other areas (2003, 195).

In fact, most cells in the visual cortex will respond to a range of stimuli including a moving line, stationary lines of a particular orientation, a range of wavelengths as well as achromatic stimuli and their response will also be modulated by stereoscopic factors (Schiller 1996, 26-7).

Although these empirical challenges to the single cell studies are important, they also face a conceptual problem: they may not be studying the vehicles of representation. Single-cell studies are based on the neuron doctrine (Barlow 2001); the idea that a neuron's receptive field describes the percept caused by the excitation of the neuron (Wandell 1995, 188). However, it is an open possibility that the brain uses distributed representations in that each piece of information is represented by a pattern of activity over many different neurons and each neuron contributes to the representation of many different pieces of information. If this were the case then an examination of the preferred stimuli of neuron would not provide any insight into what is being represented because individual units/neurons are not the representing vehicles.

Lehky and Sejnowski's (1988; 1990) experiments with an artificial neural network which determined shape from shading demonstrate how misleading simply examining the receptive field of an individual neuron can be. They found that units in the hidden layer of a shape from shading network had receptive fields similar to those in the visual cortex which are standardly thought to be edge detectors (Palmer 1999b, 185). But, the units were involved in calculating shape from shading, not detecting edges. This suggests that an 'examination of the receptive fields of individual units does not make apparent what the network is doing' (Lehky and Sejnowski 1988, 454). Although we still do not know precisely how to map the architecture of a network onto the architecture of the brain, it is fairly clear that the neuron doctrine is in conflict with a commitment to distributed representations. If information is distributed across many neurons, then the single-cell studies that dominate current research will be unable to provide us with information as to the representational content of the visual system. In light of this, the most sensible conclusion seems to be that the current physiological evidence is inconclusive.

The psychophysical results are also not as simple as Livingstone and Hubel's model would require. Numerous studies have demonstrated that motion perception is possible at isoluminance

(Cavanagh and Anstis 1991; Cavanagh et al. 1985; Cropper 2005; Gegenfurtner and Hawken 1995; Papthomas et al. 1991; Zeki 1993) and more recent studies of depth perception at isoluminance also suggest that it is possible to use colour information to perceive depth (Kingdom and Simmons 1996; Logothetis et al. 1990; Scharff and Geisler 1992; Simmons and Kingdom 1997). In general it seems that the current evidence is inconclusive, but that it is consistent with the claim that there is an interdependence between the representation of spectral (particularly wavelength) spatial and figural properties.

It should be noted that even if it were demonstrated that there is functional segregation in the visual cortex, this finding could be consistent with the claim that various visual features are modelled together. There is no reason why there cannot be separate parts of the overall representational model that are specialized for modelling different things. All that is required is that there are sufficient connections between the separate areas such that the way in which spatial properties are represented influences the way spectral properties are represented. In other words, the current view of vision is consistent with there being numerous interconnected networks in the visual system. In line with this picture there is anatomical evidence that there are numerous bi-directional connections between the different areas of the visual cortex (Lennie 1998 provides a good review of this literature).

The claim that the ways in which spectral, spatial and figural properties are represented are co-dependent is consistent with our current empirical knowledge. Moreover, because the nature of the light reaching the eye is determined by all of these factors, it makes sense to represent them together. But, the idea that the purpose of *colour* vision is to assist us in the general task of using information in the ambient light to guide behaviour, is not particularly new (see for instance Akins 1996a; 2001; Akins and Hahn 2000; Milner and Goodale 1995; Mollon 1989; Thompson 1995). More importantly, we have not yet determined the nature of the facts that are represented in colour experiences. Do colour experiences represent surface

reflectance or do they represent the relations between the spectral, spatial figural properties of the world? What we need is a *representational* account of this co-dependence.

2.3 Modelling the Visual World

We have seen that the appearance of a surface depends not only on the reflectance of the surface but also on the spectral, spatial and figural nature of the entire visual field. The dependence also goes the other way. The way in which the spatial properties of a scene are represented depends on the spectral properties of the stimulus. We have also seen that this co-dependence is part of the visual system's constancy mechanism. In order to more accurately represent the visual world the visual system models the spectral, spatial and figural properties of the stimuli together. Although we can make sense of the co-dependence between the representations of spatial, figural and spectral properties, we have yet to determine the represented object of colour experiences. In this section I will argue that adopting a molecular scheme of representation provides us with a new way of thinking about the content of visual experiences. In particular we will see that it is possible to make sense of the interdependence between the representation of spectral, spatial and figural information while still maintaining that colour experiences are representations.

2.3.1 Molecular and Lexical Schemes of Representation

Cummins (1996) draws a distinction between molecular and lexical schemes of representation. A molecular scheme of representation is one in which the minimal semantic elements "have no meanings independent of their occurrence in a containing "complete" representation" (Cummins 1996, 76). In contrast, in a lexical scheme, the minimal semantic elements have meaning independently of their context in the larger representation.

Consider two different schemes for constructing the simple picture in figure 2.5. The elements in the boxes on either side of the picture are the basic representational building blocks out of which the picture is constructed. We can see that for the molecular scheme on the left, the basic representational building blocks have their meaning partially[27] in virtue of their place in the larger representation. For example the circle represents a head or a wheel partly in virtue of its place in the larger representation. In contrast the representational building blocks in the lexical scheme on the right do have a meaning independently of their place in the larger representation. The house represents a house independently of its relations to the car and the person.

Another example of a representational system that is molecular is a mercury thermometer. It represents temperature because the relations between variations in the volume of mercury resemble the relations between variations in the temperature of the substance that the thermometer is in contact with. For example if volume x is greater than volume y then the temperature which corresponds to x is hotter than the temperature that corresponds to y. Similarly if volume b is in between volumes a and c, then the temperature that corresponds to b is in between the temperatures that correspond to a and c. However, like the circle in figure 2.5, an individual volume of mercury does not represent temperature. Rather, it is (partially) in virtue of its relations to the other volumes of mercury that a particular volume represents a particular temperature. In general, for molecular schemes individual representing vehicles have content partially in virtue of their relations to other vehicles, whereas in lexical schemes vehicles have content in and of themselves. [28]

[27] In this instance they also have their meaning partially in virtue of their shape.

[28] It is important to note that molecular schemes are not necessarily holistic in any problematic way (Cummins 1996). "It is diagnostic of holistic schemes that changing the meaning of existing basic elements (or adding or removing them) changes the meaning of others" (Cummins 1996, 75). The problem with holism is that it is plausible that the basic representational elements of two people are different such that every representation they had would mean different things. This would have the effect that intentional generalizations and therefore a scientific study of the mind would be impossible (Fodor and Lepore 1992, 15). However it is not the case that adding new basic representational elements changes the meaning of all of the

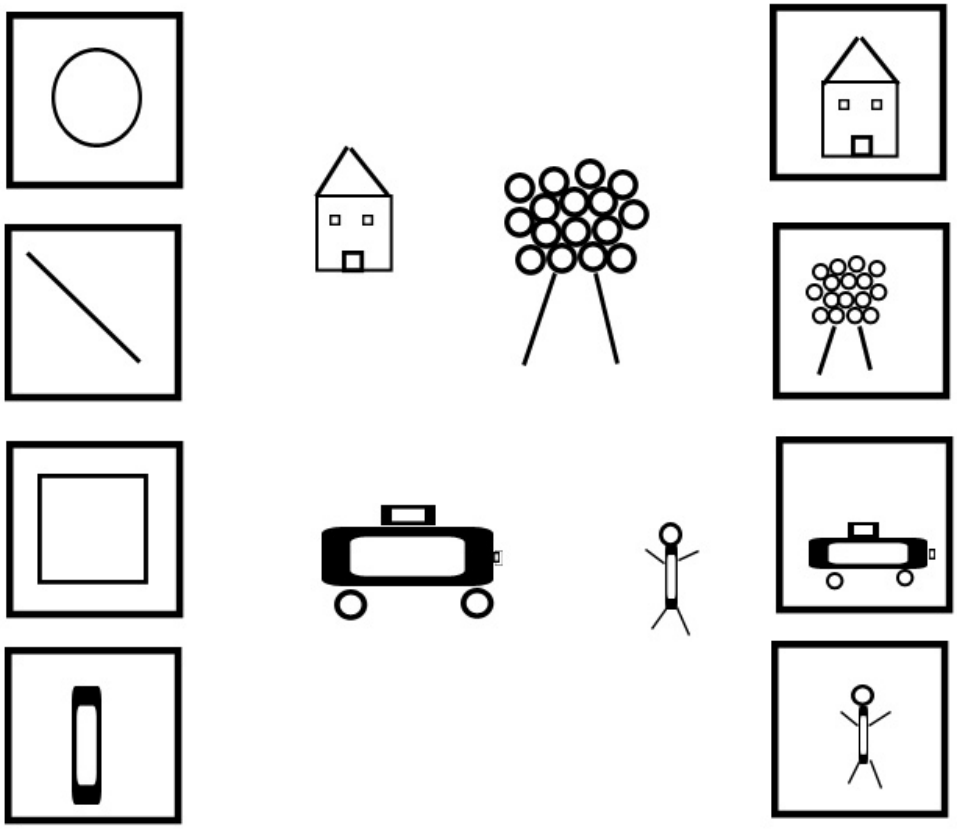


Figure 2.5: Molecular and Lexical Schemes for constructing a simple picture. The scheme on the left is molecular whereas the one on the right is lexical (after Cummins 1996, 76)

The crucial aspect of molecular schemes of representation is that, because a vehicle has its content partially in virtue of its relations to the other vehicles, it is a mistake to ask for molecular schemes what an individual vehicle represents in and of itself. For example the circle

existing elements for molecular schemes. For example adding a new shape to the representational palette in figure 5, say a star, doesn't affect the fact that the circle represents a head or a wheel. In general for molecular schemes, although content is determined by the relations between the vehicles, adding or removing a basic element does not necessarily change the meaning of *all* of the vehicles. This means that it is possible for two representational systems to mean the same thing by a representation without having to be representationally identical. Molecular schemes of representation are not necessarily holistic, at least not in the sense that threatens the possibility of intentional explanations.

represents nothing in and of itself. It is only when it is placed in a larger representational context that it represents a head or a wheel or some leaves. Similarly, a particular volume of mercury considered in isolation, does not represent temperature. It only represents the temperature of the substance it is in contact with in virtue of its relations to the other volumes of mercury. In general, for molecular schemes of representation, we can only talk about the content of an individual vehicle in virtue of its place in the overall system of vehicles. It is primarily the systems of vehicles that represents, such that an individual vehicle only has meaning because of its place in the overall set of vehicles. If colour vision were molecular, instead of focusing on the question "what do red experiences represent?" we would need to ask, "what does the system of representing vehicles of which red experiences are a part of represent?"

2.3.2 The Molecular Account of Vision

This change in focus, from individual vehicles to the system of vehicles, makes it possible to develop an account of the facts represented in colour visual experiences that can do justice to the psychophysical data. Given that a consideration of spatial properties is useful in more accurately representing the spectral properties of the world, and vice versa, perhaps these properties are represented together in one large representational model. Spectral and spatial information interact in determining the nature of the light reaching the retina. By allowing these features to interact in the visual model of the world, a more accurate representation of the spectral and spatial properties of the world can be achieved.

The idea is that the complex representing vehicle of which colour experiences are a part represents not only the spectral properties of the world, but also the spatial and figural properties of the world.[29]

[29] The current suggestion is similar to Akins (2001) suggestion that the primary purpose of the spectral system may not be to see objects as coloured, but rather to assist us in the general task of sensorimotor guidance using visual information.

Within this visual model, colour experiences may represent the spectral properties of objects. But, they would only have this representational role derivatively. The representational relation between phenomenal redness and a particular surface reflectance would be dependent on its place in the larger representing vehicle. As the larger vehicle also represents some non-spectral features of the world, the appearance of a surface depends on these non-spectral features.

Importantly, adopting a molecular scheme of representation provides a *representational* way of implementing this idea. Recall that one of the major difficulties that the psychophysical data poses for a representational account of colour is that there does not seem to be any simple mapping from red experiences to some objective property of surfaces – "redness" – in the world. This raises the possibility that red experiences do not represent anything at all. But, if colour vision is molecular, then it is a mistake to look for a mapping from an individual experience to some feature in the world. Individual colour experiences would not represent in and of themselves. There is no simple mapping from the circle in figure 2.5 to objective properties. In one context it represents a head, in another a wheel and in another leaves. The fact that the circle can represent many different things doesn't mean it doesn't represent anything at all. Rather, it is simply a result of the fact that it represents partially in virtue of its relations to other things. Similarly, red experiences can represent, even though there is no mapping from red experiences to some objective property. In one instance a particular red experience may represent a surface with one particular spectral reflectance and in another instance the same type of experience may represent a surface with an entirely different SSR. This is no problem; what represents is the larger representing vehicle, of which red experiences are a part. If that larger vehicle can be about more than one thing, or if the experience can be a part of more than one larger representing vehicle, then the experience can represent more than one thing. As with the circle, the absence of a simple mapping from red experiences to objective properties does not mean

that they do not represent anything. Rather it is a reflection of the fact that red experiences do not represent in and of themselves.

The second difficulty posed by the psychophysical data was reconciling the data with the introspective view of colour. It seems that our colour experiences are about the material properties of surfaces. But, the psychophysical data suggests that colour experiences depend on the spectral, spatial and figural properties of the entire visual scene. The puzzle is: how can we be so wrong about our experiences? On the molecular view of vision the puzzle is solved. This is because there is a sense in which a particular colour experience, at a particular time, can represent a specific surface. It is just that the same experience, in a different context, can represent something very different, because what the experience represents is influenced by its context.

Colour experiences are (often) used to represent the nature of a particular surface. Nevertheless, if the spectral, spatial or figural nature of the visual scene changed, the same surface, under the same illumination, can be represented differently. On the molecular account of vision, this is no mystery. Consider again the circles in figure 2.5. Although in every instance what the circle represents depends on its relation to the other shapes, it is nevertheless the case that in any individual instance, the circle represents something specific, such as a head or a wheel. Similarly, within the context of a visual experience, an individual colour experience may represent something specific about a surface. It is just that it only represents the surface in virtue of its relations to the other experiences. If these change, so too can what the red experience represents change.

Consider for example the image of the box of mints and the bowl in figure 2.6. Examine the reflection on the bowl and the highlight on the edge of the box of mints as indicated by the two arrows. These highlights look the same, yet the real-world stimuli that produced these two regions have very different properties. The region indicated by the lower arrow is the result of a surface that reflects predominantly long-wavelength light being illuminated by an illuminant with wide distribution of photons. The region indicated by the upper arrow is

produced by a surface that reflects light of all wavelengths being illuminated by an illuminant with a predominance of long-wavelength photons. There are two completely different sources, yet two (relatively) identical appearances.

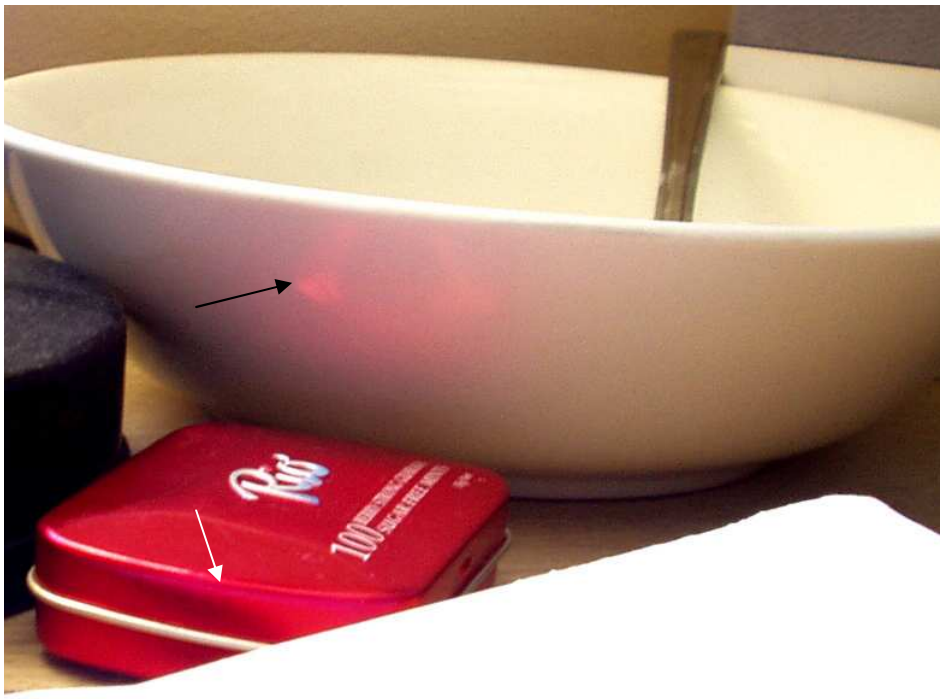


Figure 2.6: The two regions indicated by the arrows appear to be the same even though quite different combinations of surfaces and illuminants produce them.

The point is that we cannot ask what that type of red experience represents independently of its place in the larger visual scene. In one place it represents a white surface under red illumination; in another it represents a red surface under white illumination. However, within the context of a larger visual scene, we can specify what the red experience represents. In one instance it represents a surface that reflects predominantly long-wavelength light being illuminated by an illuminant with wide distribution of photons. In the other instance it represents a surface that reflects light of all wavelengths being

illuminated by predominantly long-wavelength photons. However, take away the larger context, and you have taken away the mapping from the experience to the objective arrangement of surfaces and illuminants. The red regions only map onto the objective properties that they do in the context of the larger representation.

One is asking the wrong question when one asks what red experiences represent. It is more accurate to ask: what do colour visual[30] experiences represent? The answer is they represent, at least, the relations between the spectral, spatial and figural properties of the world. So, a simple colour visual experience may represent not only the relation between the reflectance of the region and its surround, but also the spatial relations of the surfaces, such as the fact that the surface is part of a particular object and spatially distinct from its surround.

A more realistic example would be the image in figure 2.6. In this example shading information helps us to determine that the bowl and the box are both curved, that the bowl is concave and that the illumination is coming from above and behind the photographer. We can also gain information about the textures of the surfaces such as that both the bowl and the box are shiny and relatively smooth. In fact, the task of trying to list what the colour aspects of the image do and do not represent quickly demonstrates that the amount of information in any one colour visual experience is enormous. It does not seem possible to carve up this information and attribute it to different aspects of the experience. Rather, the spatial, spectral and figural aspects of our phenomenology work together to represent the spatial, spectral and figural aspects of the visible world.

In general, within the context of a larger visual representation it is possible to get a mapping from a particular colour experience to a particular surface property. But, because this representational mapping holds in virtue of the place of the particular experience in the larger

[30] Note that I am using “colour visual experiences” to refer to the visual experiences of a normal human being that include colour phenomenology. In particular I want to distinguish colour visual experiences from achromatic experiences (of either the normal human being or the achromat). Also note that I am not claiming that achromatic human visual experiences cannot represent spatial, spectral and figural properties. Rather the claim is simply that we can often do it better with colour.

visual experience, and because the larger visual context can change, the one type of phenomenal colour can be used to represent more than one type of feature in the world. It is not possible to consider the representational content of a particular type of phenomenal colour in isolation, because it only has its representational content in virtue of its relations to the other phenomenal qualities that make up the experience. There is a limited sense in which an individual colour experience may be about the spectral properties of a surface. Within the context of a larger visual representation, it is possible to get a mapping from a particular phenomenal colour to a particular surface feature. Importantly, this sense in which colour experiences in a particular representational context are about particular surface seems sufficient to respect our colour phenomenology. However as it is primarily the colour visual experience as a whole that represents, and because colour vision models some of the non-spectral properties of the world as well as some of the spectral features, it is more accurate to say that colour experiences are *part* of our visual model of the world and that this visual model represents (at least) the spectral, spatial and figural properties of the world.

On the molecular view of vision the fact that the spectral, spatial and figural properties of the world are represented together becomes comprehensible. The nature of the retinal image is determined by the relations between the spatial, spectral and figural features of the external world. So, it makes sense for the visual system to model these features together. Instead of being a complication that needs to be explained away, the interdependence between the representation of spectral and spatial information is evidence of the sophistication of the visual system. More importantly we can provide a *representational* account of how our visual system is able to achieve this task. In molecular schemes of representation, the fact that there is no mapping from individual representing vehicles to some objective property does not mean that the vehicles have no representational content. Rather, it is a consequence of the fact that individual vehicles do not represent anything in and of themselves. Adopting a molecular scheme of

representation makes possible a sophisticated account of the represented object of colour visual experiences that can do justice to the psychophysical data.

2.4 The Colour Vision Facts

We have seen that colour vision is molecular, such that there is no simple answer to the question: "what do colour *experiences* represent?" Nevertheless, if we are to demonstrate that there are *physical* facts that we can represent only in colour experience, we need to get a clearer understanding of the nature of the facts that are represented in colour experience. In this section I will home in on the nature of the colour visual facts, that is, the facts that are represented in (normal human) colour vision.

One might be tempted to think that it is now necessary to consider the difference in the represented objects of colour vision and non-colour vision. That is, it might seem that it is necessary to home in on the specific representational work that is being done by colour experiences. The problem with this is that there are no general easy answers to this question on the molecular account of colour vision developed above.

Fortunately it is not necessary to home in on the specific work being done by colour phenomenology. Colours are just the salient example chosen by Jackson in order to pump the intuition that the only way to know what an experience is like is to have the experience. As became clear in our consideration of Mandik's (2001) reply, the real contrast is that between the (putative) facts that we can represent only by having experiences (colour or otherwise) and those facts (such as scientific facts) that can be represented without having any specific type of experience. Remember the worry is that it doesn't seem to matter what sorts of experiences you can have when it comes to knowing the scientific/physical facts, whereas it does seem to matter when it comes to knowing what an experience is like. The important

epistemic contrast is not between colour vision and non-colour vision, rather it is between colour vision (as an example of "experiential" representation) and non-experiential, non-visual knowledge. The upshot is that it is not necessary to try and pinpoint the representational work done by colour experiences. We can focus on colour vision in general instead. In order to make clear that this shift in focus has occurred, instead of talking about the 'colour facts', I will now focus on the 'colour visual facts', that is, the facts that are represented in colour vision, as a whole.

What are the facts that are represented in colour vision? This is not an easy question to answer in general. A complication that arises from the suggestion that colour vision is molecular is that it is very difficult to give a detailed but general characterisation of the nature of the facts that are represented in normal human colour vision. Just as there are so many different pictures that can be built out of the shapes in the molecular scheme of representation in figure 2.5, so too are there many different complex visual representations that the normal human visual system can construct. Given the nature of the input on the retina it is possible to home in on, amongst other things, the spectral, spatial and figural properties of the world. The psychophysical data considered above indicate that colour vision plays an important role in the representation of the spectral, spatial and figural properties of the world.[31] It is difficult to be more precise than this when talking generally. Instead, I think it is better to focus on individual examples, such as that in figure 2.7.

[31] It is still unclear what role spectral information plays in the representation of depth and motion (Cropper 2005; Heywood et al. 1998; Jimenez et al. 1997; Kooi and De Valois 1992; Watanabe and Cavanagh 1992).

Figure 2.7 A state of affairs of the sort that is represented in colour vision.

The state of affairs that is depicted in figure 2.7 is quite complex. We can get information about the relative positioning of the camera, car, sun and surrounding vegetation. We can see that the concrete on which the XJS is sitting is wet; that the surface of the XJS is highly reflective; that the bonnet of the car has a slightly raised ridge in the middle and that the sides of bonnet are raised even higher (if you are more informed you can see that it has that classic “jag” shape); that the inside of the headlights is concave; that there is more vegetation above and to the left that is being reflected from the bonnet of the car; that there are trees and a building behind and to the right of the camera that are being reflected from the side of the car at the front; that the tires not smooth but rather have a tread cut into them; and so on. We can also use the specular reflections on the bonnet of the car to gain information about the shape of the bonnet (Fleming et al. 2004). There is also a wealth of information about the spatial relations between the car, the sun, the vegetation, the concrete, and between various aspects of the car. We can also gain information about the nature of the component surfaces, such as that the front bumper bar is matt whereas the surface

of the bonnet is shiny and tends to reflect photons with a wavelength around 510nm (see Shafer and Maxwell 2000, for a discussion of the physics of surface texture). This diverse range of information concerning the spectral, spatial and figural properties of the state of affairs that was captured in the photo is an example of the type of facts that are represented in colour visual experiences.

To return to the knowledge argument this means that what is represented in colour visual experiences are complex states of affairs concerning the relations between the spectral, spatial and figural features of the world. What is not clear, however, is that these complex states of affairs are distinct *facts* over and above the spectral, spatial and figural features of the world which constitute them. If the account being provided is to be different from the old fact reply, then the complex states of affairs that are represented in colour visual experiences need to be metaphysically distinct from their constituents. If we cannot show this then we cannot rule out the possibility that colour vision simply provides a different way of representing the same facts about the spectral, spatial and figural properties of the world that are represented in other ways. I will turn to examine this issue in the next chapter.

Chapter Three

The Emergence of the Colour Visual Facts

We have seen that colour visual experiences represent complex states of affairs concerning the co-dependence of the spectral, spatial and figural properties of the world. In having colour visual experiences, we gain a more detailed model of the visual world. In this chapter I will consider whether the complex states of affairs that are represented in colour experience are distinct facts over and above the various phenomena that constitute them. We have seen that colour vision represents quite complex phenomena. However, it is far from clear that what we represent in colour vision is not just a gerrymandered collection of things. What makes the state of affairs depicted in figure 2.7 *new* and distinct from the various elements (such as the car and the buildings and the light) that constitute it? The problem that we now face is the one that caused problems for Mandik (2001). Just showing that vision represents complex states of affairs does not imply that these complex states of affairs are distinct from the elements that constitute them. We need to not only account for the represented object of colour vision; we also need to clearly show that what is represented is a *new* and distinct fact in order to avoid providing another version of the old fact reply.

Before this specific question concerning the metaphysical nature of the colour visual facts can be answered it is necessary to develop a more general metaphysics. Because we are interested in the nature of the relation between the complex states of affairs that are the colour visual facts, and their constituents, the focus will be mereology; that is, on the relation between parts and wholes in our world. I will argue that a proper examination of what science has already told us about the world clearly demonstrates that there are emergent wholes which are metaphysically distinct from their constituents.[32]

This chapter will begin with an examination of the many conceptions of emergence that have been developed. I will argue that there are two core ideas: that emergent phenomena are novel with respect to their constituents; and that emergent phenomena are inexplicable in mereological terms. Next, I will turn to consider some scientific examples to see whether we find mereological novelty or inexplicability. We will see that although there are many examples of mereologically novel phenomena, we have mereological explanations of them. In the third section I will develop a metaphysics on which there is explicable, mereological, novelty. We will see that, unlike the standard form of non-reductive materialism, which defines metaphysical levels in terms of multiple realisability, the "structuralist" metaphysics developed in this chapter can provide for genuinely emergent causal powers. We will see that the key is to base the novelty of a complex phenomenon in the *physical organisation* of its parts and thereby deny any identity of the emergent whole with its constituents. Finally, I will argue that the states of affairs that are represented in colour visual experiences are structurally emergent and therefore are metaphysically distinct from their constituents.

[32] Much of the material in this chapter is the result of collaborative work with Jon Opie. I am particularly indebted to him for suggesting this way of construing mereological emergence and for pointing me in the direction of Bunge's (1977; 1979) and Auyang's (1998; 2000) work.

3.1 Emergence

3.1.1 The General Notion

There are almost as many notions of emergence and ways of distinguishing between the different notions of emergence as there are people writing on this topic (Alexander 1920; Auyang 1998; 2000; Bechtel 2001; Bechtel and Richardson 1992; Broad 1925; Bunge 1977; Cunningham 2001; Humphreys 1997; Kim 1999; McLaughlin 1992; O'Connor, T. 1994; Pereboom 2002; Pereboom and Kornblith 1991; Ramsey 1997; Schroder 1998; Van Brakel 1997; Wimsatt 2000). Despite running the risk of adding to this confusion, it is necessary to enter the quagmire to try and survey the lay of the land. In this section I will sketch some of the standard notions of emergence and try to find some unity in the diversity.

Emergentism is a position that is motivated by a need to account for the complexity we see in our world. When we look around the world, we are faced with two, apparently contradictory, claims. On the one hand we want to accept some minimal form of materialism. Science seems to be telling us that everything is made up of those things that physicists talk about. In particular it seems that everything is composed of the entities of microphysics, such as quarks or strings. It seems that if you took away the entities of microphysics from the universe, there would be nothing left. On the other hand, there are complex things in the world, such as chairs, people and societies, which in some sense seem to be more than the entities of microphysics. For example, the table on which my computer is resting does not seem to be just a pile of atoms; it is a pile of atoms arranged in a certain way. With the aid of a sledgehammer I could transform the same set of atoms from a table into kindling. Similarly, what makes something a chair seems to have less to do with what it is made of and much more to do with what use it can be put to – we use it to sit on. How are we to understand the relation between the entities of microphysics and the “higher-level”

things we see around us? Is everything nothing but the entities of microphysics? Or are higher-level things somehow distinct from their microphysics parts? Will all theories be reduced to microphysics theories? Or are theories of higher-level things somehow independent of theories in microphysics?

One response to this conundrum is to adopt a form of radical eliminativism and hold that the complexity is apparent, and that really everything is nothing but the entities of microphysics. At the other end of the spectrum, one can adopt a form of substance dualism and hold that the only way to account for these apparently new, complex phenomena is to propose the existence of new fundamental entities, such as an immortal soul.[33] Emergentism attempts to steer a middle course between substance dualism and eliminativism. The emergentist holds that the novelty is not a result of new basic things; rather, it is somehow the result of arranging the basic entities of physics in the right way. The novelty emerges out of the basic entities of microphysics, but is somehow more than these basic entities.

The literature on emergence is so diverse that there is not even a single answer to the question: what is emergent? Usually “emergence” is used to characterise a relation between a complex whole and its constituents (Auyang 1998; Broad 1925; Bunge 1979; Kim 1992; McLaughlin 1992; O'Connor, Timothy and Wong 2002), but in the literature on dynamic systems “emergence” can be used to characterise a relation between two different states of a complex system where later states are somehow more than their predecessors (Auyang 2000) . Given that out ultimate argumentative aim is to show that the complex states of affairs that are the colour visual facts are metaphysically distinct from their components, the following discussion will focus on emergence as a relation between parts and wholes. It should be noted from the beginning that such a mereological account

[33] A third alternative is some form of property dualism. Distinguishing between emergentism and property dualism is no small task. I will take up this issue in the following chapter, once I have developed the general account of emergence.

of emergence might not provide a universal account of emergence, as it need not include dynamical forms of emergence.

Even within the literature that focuses on emergence as a mereological phenomenon, there is still no agreement on what is emergent. In particular, there are mereological accounts of the emergence of both things and properties. It has been suggested that the emergence of properties may be fundamental and that a thing is emergent if it has at least one emergent property (McLaughlin 1992; Stephan 1992). However, it is not clear that there can be a property without there being a thing that has the property. The existence of a property may be dependent on the existence of a thing, such that we cannot say that it is primarily properties that emerge. As McLaughlin (1992) and Stephan (1992) acknowledge, dealing with these questions requires us to settle some deep metaphysical issues that are far beyond the scope of this discussion. In order to avoid such issues I will talk about the emergence of both things and properties, depending in large part on the nature of the example being considered. Whether it turns out that the emergence of properties is dependent on the emergence of things, or vice versa, is an issue for another day. When talking more generally I will use the phrase emergent “phenomena” as a purposely vague term that can include both properties and things.

There are a number of different marks or conceptions of emergence. Often these are not intended to be taken as definitions of emergence, nor are they viewed as constitutive of emergence, rather some of these are taken to be more symptoms of emergence. The distinction between a mark of emergence and something that is constitutive of emergence is not a sharp one. What is a mark for one person can be taken to be constitutive of emergence for someone else. As it is not possible to determine whether something is a symptom of emergence or constitutive of emergence independently of a general account of emergence, I will not attempt such a categorisation. Nevertheless it is important to appreciate that not all of the notions considered below are thought by everyone to be constitutive of emergence: some think they are consequences of emergence. It

should also be noted that most people present a package deal by combining a number of these marks together to form an overall conception of emergence. The number of potential views is much larger than the number of individual components. We can nevertheless find some order in the diversity. In particular, it is possible to make a basic division between metaphysical and epistemological notions. The former are concerned with the emergence of phenomena themselves whereas the latter are concerned with the nature of our knowledge or theories of those phenomena.

3.1.2 Epistemic Notions of Emergence

There are three epistemic notions of emergence: non-deducibility; unpredictability; and inexplicability. Broad (1925), and Hempel and Oppenheim (1948) have claimed that one mark of emergence is the inability to *deduce* the nature of the system from knowledge of the nature of its parts. Broad suggests that:

> there are certain wholes, composed (say) of constituents A, B, and C in a relation R to each other; all wholes composed of constituents as the same kind as A, B, and C in relations of the same kind as R have certain characteristic properties; A, B, and C are capable of occurring in other kinds of complex where the relation is not of the same kinds as R; and the characteristic properties of the whole r(A,B,C) cannot, even in theory, be deduced from the most complete knowledge of the properties of A, B, and C in isolation or in other wholes which are not of the form R(A,B,C). (Broad 1925, 61)

In contrast, Nagel (1961), Popper and Eccles (1977) and Broad (1925) have suggested that a mark of emergence is the *unpredictability* of an emergent phenomenon given knowledge of its parts. The idea is that complete knowledge of constituents will be insufficient to predict the occurrence or nature of an emergent phenomenon. For example, it does not matter how much you know about physics, if life were

emergent, then it would be impossible to predict the occurrence of living things based solely on knowledge of their physics parts. Finally, Alexander (1920), Kim (1999) and Schroder (1998) have suggested that emergent phenomena are inexplicable in mereological terms. For example Alexander writes:

> The existence of emergent qualities thus described is something to be noted, as some would say, under the compulsion of brute empirical fact, or, as I should prefer to say in less harsh terms, to be accepted with the "natural piety" of the investigator. It admits no explanation (Alexander 1920, 47).

Note that this is an in principle claim. It is not that we have not yet explained emergent phenomena. Rather the claim is that emergent phenomena can never be explained in mereological terms; the emergence of complex phenomena is a brute fact that needs to be accepted with "natural piety".

What the notions of unpredictability, non-deducibility and inexplicability have in common is that they all concern our ability to *know* about the nature of emergent phenomena based purely on knowledge of their constituents. The claim is that we cannot predict, deduce or explain the nature of the whole from knowledge of the constituents. I suggest that the most fundamental epistemic notion is inexplicability.

Hopefully it is clear that non-deducibility is simply inexplicability as understood within the framework that holds that explanation is deduction. According to a prominent account of scientific explanation, to give a scientific explanation of something is to deduce its occurrence from a covering law and background conditions (Hempel and Oppenheim 1948). On such an account, to explain the nature of a complex thing in terms of its constituents, it would be necessary to deduce the nature of the whole simply from knowledge of the parts. If such a deduction proved to be impossible, then the whole would be inexplicable in terms of its parts. As it is controversial to claim that scientific explanation is deduction from a covering law (Kitcher 1989;

Salmon 1984; Woodward 2003), it is far from clear that non-deducibility prevents explanation. Thus it is best to take inexplicability as the primary notion and view non-deducibility as one, theory-laden, articulation of inexplicability.

Similarly, it is no longer clear that it is best to understand the epistemic issues surrounding emergence in terms of unpredictability. In the classical account of science, which views explanation as deduction, prediction is also deductive (Popper 1963; Whewell 1857). In fact, on this view, the difference between explanation and prediction is mainly whether the observation of the event has been made (in which case we deduce the occurrence of a phenomenon in order to explain why it occurred) or not (in which case we deduce its occurrence as a prediction). This idea that we deduce predictions from our theory in order to test the theory has also been questioned (Kuhn 1962). Thus I suggest that, as with deducibility, it is best to view predictability as a theory-laden attempt to get at the more fundamental epistemic notion, namely inexplicability. Presumably the reason why we cannot predict the occurrence of an emergent phenomenon based on knowledge of its parts is because it is inexplicable in mereological terms. Thus, as with non-deducibility, it seems that inexplicability is a more fundamental notion that unpredictability.

We can therefore see that one of the themes running through many people's conception of emergence is that mereologically emergent phenomena are inexplicable in mereological terms. The suggestion is that we cannot explain the nature of the whole in terms of its parts. It is important to note that many people who hold that emergence involves inexplicability often hold that inexplicability is not constitutive of emergence (Broad 1925; McLaughlin 1992; O'Connor, Timothy and Wong 2002), but rather is a consequence of the metaphysics.

3.1.3 Metaphysical Notions of Emergence

There have been three main metaphysical conceptions of emergence. The oldest metaphysical notion, which stretches back before the term “emergence” was even coined, is the idea that emergence involves the non-additivity of the parts in determining the nature of the whole. It is this idea that underlies one of the more well-known slogans of emergence, namely that “the whole is more than the sum of the parts”. This idea comes from Mill (1879), who distinguished between two ways in which various causal factors can combine to produce an effect. For some causal interactions Mill suggested that ‘the joint effect of several causes is identical with the sum of their separate effects’ (Mill 1879, 428). The paradigmatic example of such causal interactions is the composition of forces. The total force acting on a body is simply the sum of the component forces, as calculated via vector addition. If you are being pushed from behind and the left then you will move forwards and to the right. But, Mill suggested, various causal factors do not always combine in an additive way. For example he suggested that

> The chemical combination of two substances produces, as is well known, a third substance with properties different from those of either of the two substances separately, or of both of them taken together. Not a trace of the properties of hydrogen or oxygen is observable in those of their compound, water (Mill 1879, 428).

Water does not in any way seem to be the sum of hydrogen and oxygen. The combination of hydrogen and oxygen produces something entirely new that is not in any way the sum of its parts.

Emergence has also been conceived of simply as novelty of the whole with respect to its parts (Alexander 1920; Auyang 1998; 2000; Bunge 1977; 1979; Wimsatt 1994; 1997). For example Alexander suggests that

> The emergence of a new quality from any level of existence means that at that level there comes into being a certain constellation or collocation of the motions belonging

> to that level, and this collocation possesses a new quality distinctive of the higher complex. The quality and the constellation to which it belongs are at once new and expressible without residue in terms of the processes proper to the level from which they emerge (1920, 45)

Call such novelty of the whole with respect to its parts “mereological novelty”.

Conceiving of emergence as mereological novelty immediately raises the question of what makes something metaphysically novel. On what basis would we want to say that the whole is novel and metaphysically distinct from its parts? In fact, on the basis of what would we want to add *anything* to our ontology? Kim’s suggestion that to be real is to have causal powers is prima facie plausible:

> For something to be real is for it to posses causal powers (let us call this “Alexander’s Dictum”); mere epiphenomena have no causal work to do, and their existence makes absolutely no difference to the rest of what exists – they might as well be “abolished” (Kim 1992, 134)

This principle, named after one of the more prominent British Emergentists, certainly seems to capture a sufficient condition for entry into our ontology. There seems to be no reason to add something to our ontology that is epiphenomenal. If something makes no impact on anything whatsoever, then the most sensible conclusion is that it does not exist.

Alexander’s dictum fits well with both common sense and scientific methodology. If I suggested that there really are fairies that the bottom of my garden, but that no-one, myself included, can see, touch, hear, or detect them in any way whatsoever, then the reasonable conclusion would be that I was mistaken (and perhaps crazy) and there are in fact no fairies at the bottom of my garden. The observation and experimental manipulation of the world that are at the heart of the scientific methodology also seem to involve a commitment to Alexander’s dictum. For example, geologists ultimately admitted that continental drift occurs because there was a complex pattern of effects

that could best be explained by the hypothesis that the continents are moving (Kitts 1977; Laudan 1978; Marvin 1973). Because of the match with both every day and scientific practice it seems plausible to accept Alexander's dictum as a sufficient condition for existence.[34]

Given that we have accepted that to be real is to have causal powers, it seems that a mark of ontological difference is having a different type of causal power. That is, we could say that a complex whole is distinct from its parts if the whole has a different type of or collection of causal power(s) to those had by its parts. Note that it is important that the difference in causal powers be qualitative, not quantitative (Auyang 1998). For example, it seems appropriate to say that although a grain of sand and a pile of sand have different causal powers (only the latter could move a balance beam that has an apple on the other end); they nevertheless have the property of having a mass in common. This difference in causal powers is quantitative not qualitative. They have the same type of causal power, just different magnitudes of it. It seems that it is only if the complex whole can do a distinctive type of thing that we are warranted in concluding that it is ontologically distinct from its parts. The second metaphysical notion of emergence is that emergence is novelty of the whole with respect to its constituents. It also seems that possession of a mereologically unique type of causal power is sufficient for the whole to be mereologically novel with respect to its parts.

The third metaphysical conception of emergence also concerns causation. In particular, emergence is sometimes conceived of in terms of downwards causation (Alexander 1920; Kim 1992; 1999; Schroder 1998). If some form of mereological emergence is correct, then we will

[34] Note that it is one thing to find a causal power; it is another to attribute it to a particular phenomenon. In order to determine what is responsible for a causal power it may be necessary to adopt a condition that is more complex than simply the idea that to be real is to have causal powers. In particular, it may be necessary to go for something more like Wimsatt's notion of robustness: 'things are robust if they are accessible (detectable, measurable, derivable, producible, or the like) in a variety of independent ways' (Wimsatt 1994, 210-11). All of the examples I will be dealing with here (including the colour visual facts) are part of very well established scientific theories, and presumably have achieved this status because they are robust. Thus, for the sake of simplicity I will stick with the simpler notion, namely Alexander's dictum.

have a layered metaphysics, with emergent phenomena at one level, and their constituents at a lower level. For example, water may be at one level and its atomic constituents, hydrogen and oxygen at a lower level. Within such a metaphysics, downwards causation occurs when something from one level is able to have a causal impact on something on a lower level. If water, as an emergent whole, can have a causal effect on individual atoms, then we would have an instance of downwards causation (see Kim 1999; Schroder 1998 for a discussion of different notions of downwards causation).

I suggest that novelty, particularly when it is understood in causal terms, is a more fundamental notion than downwards causation. We have just seen that it is plausible that a sufficient mark of existence is possession of a type of causal power that is distinct from those had by other things. In particular, we have seen that if a complex phenomenon can do qualitatively different things to its constituents then we are warranted in saying it is metaphysically distinct from its constituents. The ability to have a downwards causal impact appears to be one way in which something could have a distinct type of causal power. But, the ability to have a downwards causal impact is not the only way in which something can have causal powers that are distinct from those of its parts.[35] Something could also have a causal impact "upwards" to phenomena at a higher level, or "horizontally" to phenomena at the same level. The point is that being capable of downwards causation is only one way in which a phenomenon can have a type of causal power that is distinct from those possessed by its parts. So, possession of a different type of causal power to other things is the more general notion than the ability to have a downwards causal impact.

We can also appreciate that novelty is more primary than non-additivity. The problem with the claim that non-additivity is constitutive of emergence is that there seem to be many nonlinearities that the

[35] Kim (1992) has argued that for a complex phenomenon to have causal powers that are distinct from its parts it must be able to have a downwards causal impact. This argument, however, is a variation on his general claim that emergent phenomena do not have causal powers that are distinct from those of their parts. See section 3.3 for a reply to Kim.

eliminativist can countenance. For example, as McLaughlin (1992) points out, Einstein's field equations are non-linear. The gravitational field of two objects is in no way the sum of the individual fields. However, to point to the nonlinearity of Einstein's field equations in order to support emergentism seems to provide a hollow victory for the emergentist. There seems to be no reason why the eliminativist would have a problem with such nonlinearities, because it seems that such non-linear behaviour could all be carried out by the entities of microphysics. It thus seems sensible to say that the notion that a failure of additivity may be sufficient for emergence is outdated and that the more primary metaphysical notion is that emergent phenomena are novel in comparison to their parts.

It therefore appears that there are two basic themes to the discussion of emergence. One idea is that emergent phenomena are inexplicable in mereological terms. The other is that emergent phenomena are mereologically novel and ontologically distinct from their parts, as evidenced by the possession of a different type of causal power to their parts. Other conceptions of emergence seem to be variations on these two basic themes.

Before moving on to consider some examples I need to introduce some terminology. Emergent things are contrasted with aggregates and emergent properties are contrasted with resultant properties. Mass is a paradigmatic example of non-emergent, i.e. resultant, property. A resultant property is a property of the whole that the parts also possess. For example, the mass of a pile of sand is a resultant property. Each individual grain of sand has a mass and the mass of the pile is the sum of the masses of the individual grains. Note that a pile of sand may still have causal powers that its constituents lack. For example an individual grain of sand cannot move a balance beam that has a mass of 20kg on the other end. But, the difference in causal powers is merely quantitative, not qualitative. In general, a property of a complex thing is resultant if the parts also have the same type of property.

It is quite difficult to come up with an example of an aggregate, a *thing* that does not have *any* novel characteristics in comparison to its parts. Even apparent aggregates, such as a pile of sand, have properties, such as shape, which seem to be novel with respect to its parts. The pile has a shape that is of a different type to that had by its micro-constituents. There may be principled reasons why it is difficult to find aggregates. Wimsatt suggests that it is only those properties that have a major conservation law associated with them (namely mass/energy, momentum and net charge) that are resultant (Wimsatt 1997). As most things have properties other than those for which there is a major conservation law, most things have some novel characteristics. Bunge suggests that an aggregate is a collection of items not held together by bonds, such that it lacks integrity or unity (Bunge 1979, 4). As the constituents of most things are bound together by chemical bonds, or sub-atomic forces, we should expect that there are very few things that are aggregates. It seems that if our criterion is purely mereological novelty, then emergence may be ubiquitous. This may be why some think that mereological novelty by itself provides a "weak", uncontroversial conception of emergence (Schroder 1998). However a "weak" uncontroversial form of emergence can still provide us with the mereological novelty of the whole with respect to its parts that we require.

Now that we have an understanding of the basic notion of emergence I want to turn to consider what science has to tell us about the world. In particular I will examine whether the metaphysical world-view that is motivated by a consideration of science supports either the claim that there are complex phenomena that are inexplicable in mereological terms or the claim that there are complex phenomena that are mereologically novel in that they posses causal powers that are distinct from those of their constituents.

3.2 Mereological Inexplicability and Novelty in Science

Ultimately, the success of any metaphysics is determined (at least in part) by the extent to which it fits with the world-view provided by science. In this section I want to consider what science tells us about the relations between complex "wholes" and their constituents. In particular, I will consider whether a scientific metaphysics countenances the existence of mereologically phenomena with causal powers that their parts lack. If we can find some examples of such mereologically novel phenomena, we will also consider whether they are inexplicable in mereological terms. Thus, the aim of this section is to see whether we actually find inexplicable mereological novelty within the sciences. We will see that when we look around at the sciences we find numerous examples of novel but mereologically explicable phenomena.

3.2.1 Water[36]

One of the most philosophically popular examples of an identity is that of water and H_20. Although it is undeniably correct to say that water is *composed* of H_20 molecules, when we examine the relation in more detail, it becomes apparent that the slogan "water is H_20" is an oversimplification. In particular, if we follow Alexander's dictum and take possession of distinct types of causal power as an indication that there are different phenomena, then it becomes clear that water is novel and distinct from H_20 molecules. Nevertheless, it is possible to explain the properties of water in terms of the nature of the H_20 molecule.

The term "water" is vague. First of all, just about everything we call "water" contains particles besides H_20. So, we need to talk about "pure" water. But, this is a trivial complication compared to that which

[36] This example is taken from Bunge (1979) and Auyang (1999)

arises from the fact that "water" has many different states. Does "water" refer to the liquid, solid, gaseous or crystalline versions of H_20? This is a particularly relevant question given our focus on causal powers, because "water" in these different states has different types of causal powers. Whether the "water" is liquid, solid or gaseous matters very much to the person about to leap off the diving board. Given that in ordinary usage "water" usually refers to "liquid water"[37] the most charitable reading seems to be to take "water" in the purported identity to be referring to the liquid.

The term "H_20" is as vague as the term "water". Does "H_20" refer to an individual molecule or a group of molecules (say a mole or (roughly) $6x10^{23}$ H_20 molecules)? If it refers to an individual molecule then the identity is false – an individual molecule is not a liquid. Water is not identical to an individual H_20 molecule. However, it is not simply enough to specify the number of molecules. If the molecules were evenly distributed throughout the universe they would not form any macro-substance whatsoever. Water is not identical to a collection of H_20 molecules. Even if we specify that the molecules are in close quarters, say within a region with a volume of $50cm^3$, we still have not fixed the causal powers of the system. One mole of H_20 molecules could be a gas, a liquid, a solid or a pile of snowflakes. At different temperatures and pressures the H_20 molecules will be in different states. At 1 atmosphere (atm) of pressure and 25^oC, 1 mole of H_20 molecules is a liquid. But, at 1atm of pressure and -10^0C, the same set of molecules will be a solid and at 0.01atm of pressure and 25^0C the same set of molecules will be a gas. In other words, the one set of H_20 molecules can be a liquid, solid, gas. Most importantly in these different states the H_20 molecules will have different causal powers: you can't swim in ice or drink steam.

The point of this is not to criticize those who simply talk about water being H_20 for being sloppy. I am not claiming that anyone who

[37] According to the OED definition 'water' is 'The liquid of which seas, lakes, and rivers are composed, and which falls as rain and issues from springs. When pure, it is transparent, colourless (except as seen in large quantity, when it has a blue tint), tasteless, and inodorous' (Oxford-English-Dictionary 2006).

talks about water being H_20 is ignorant of the fact that this is an oversimplification and that a more accurate identity would be something like "16ml of pure liquid water is $6x10^{23}$ H_20 molecules in close quarters at 1atm pressure and 25^oC". Rather my point is that writing the short-hand version is misleading because it masks the fact that the arrangement of the H_20 molecules is crucial in determining the nature of the system of molecules. One mole of H_20 molecules can burn or soothe a burn depending on how they are arranged. Similarly it was a group of H_20 molecules that assisted the Titanic to sail halfway across the Atlantic, and it was also a group of H_20 molecules that tore a hole in the side of the ship causing it to sink. The causal powers of a group of H_20 molecules must be dependent on more than the fact that their constituents are H_20. If not, then the differences in causal powers of groups of H_20 molecules observed above would be paradoxical. If the only determinant of the causal powers of ice and water was the nature of their constituents, then given that they have identical constituents, water and ice would be identical (Auyang 2000). This is obviously false. Water and ice have different types of causal powers. The fact that the same constituents can form different complex phenomena that have different types of causal powers means that the complex phenomena can not be identical to their constituents. There must be more to being water, snow, or ice than being composed of H_20 molecules.

Water is mereologically novel. An organised collection of H_20 molecules can have a different type of causal power to that of its constituent molecules and to those possessed by the same set of molecules in a different organisation. Nevertheless it is possible to explain the nature of water in terms of the nature of its constituents.[38] Water has a number of unusual properties. For example the liquid form of water is denser than the solid (so ice floats in water). Water also has unusually high boiling and melting points. The source of these unusual properties can be traced to the hydrogen bonds that form between H_20

[38] The details of this explanation are taken from Atkins and Beran (1989).

molecules. Hydrogen bonds are unusual because they are significantly stronger than other inter-molecular forces. This means that the molecules within water are held together more strongly than the molecules of similar substances whose molecules do not form hydrogen bonds.

The fact that ice is less dense than water can be explained in terms of the geometry of ice in comparison to water. At low temperatures, each oxygen atom is surrounded by four hydrogen atoms, two of which it is covalently bonded to and two of which it is bonded to via hydrogen bonds. The result is that the molecules form a three dimensional structure (see figure 3.1). This structure is quite rigid such that the molecules are prevented from getting close to each other. As a result of these large gaps between the molecules, the density is low. As the temperature increases, the kinetic energy of the molecules increases, on average. Some of the molecules have enough kinetic energy to pull free of the intermolecular hydrogen bonds – it is like they are jumping up and down too fast to be able to hold hands any more. As the ice begins to melt, these molecules which have broken free of the hydrogen bonded structure become trapped in the cavities of the there-dimensional structure. The result is that there are more molecules per unit volume in liquid water than in ice. So, cold water is denser than ice, such that ice floats on water.

As this example indicates, it is possible to explain the nature of water in terms of the nature of its atomic (and its subatomic) constituents. We can (roughly)[39] explain the properties of water in any of its states in terms of the properties of the H_20 molecule. This pattern, of explicable mereological novelty, can be found throughout the sciences.

[39] In fact we still lack a precise model of liquid water that can account for all of its anomalous properties. Although different models can account for different properties no model can account for them all (for more details see Chaplin 2005; Guillot 2002)

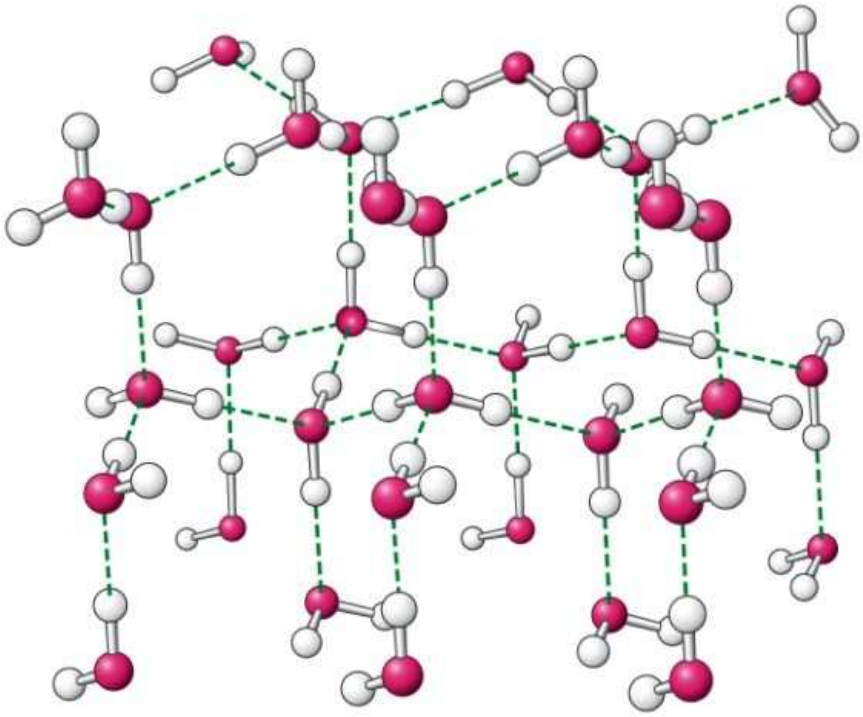


Figure 3.1 The three-dimensional structure of (hexagonal) ice. The red balls represent oxygen and the white represent hydrogen. The white lines are covalent bonds and the dashed green lines are hydrogen bonds.

http://oregonstate.edu/instruction/bb450/stryer/ch01/Slide21.jpg

3.2.2 Superconductivity[40]

Superconductors provide us with another example of explicable mereological novelty (Auyang 1998). Conductors and superconductors both have electrons that are free to move throughout the substance. Usually the electrons move in different directions such that the net current is zero. However when a voltage is applied across the substance, the electrons move in the same direction and this generates a macroscopic current. Even the best conductors have some

[40] Unless otherwise indicated, the material in this discussion of superconductivity comes from Tipler (1991). Note also that there are actually two types of superconductors. I will focus on type one superconductors as the micro-mechanism for them is relatively well understood.

resistance – if the driving force is turned off the macroscopic electric current decays and finally vanishes. This is because the electrons lose energy in the form of heat by bouncing off impurities or particles in the crystal lattice. In contrast to conductors, superconductors have a resistivity of zero – the current in a superconductor never dies out. We will see that superconductivity provides another example of explicable mereological novelty.

The first thing to appreciate about superconductors is that when in the superconductive state, the metal has a different type of causal power to that which it has at higher temperatures. At room temperatures[41] the metals that are superconductors, such as aluminium, mercury and lead, are standard conductors. That is, exactly the same set of atoms can be either a conductor or a superconductor (Auyang 1998). When electrons are placed in a cold crystal lattice they are able to do things as a group (conduct with a zero resistivity) that the same electrons, in the same crystal lattice, at a higher temperature cannot. The fact that the same set of constituents can have two completely different types of causal powers indicates that there is more to the metals possessing those causal powers than the nature of their constituents. Otherwise the difference in causal powers despite the identity of the constituents would be impossible.

Although there is more to being a superconductor than just having a particular type of microconstituent, superconductivity can nevertheless be explained in mereological terms. Superconductivity arises because of the collective action of the conducting electrons. The electrons interlock to move in concert, thereby avoiding collisions and the associated loss of energy. The basis of this interlocking is the interaction between pairs of electrons. In type one superconductors, the electrons are able to avoid collisions because they move through the lattice in pairs with one electron riding in the “slip stream” of positive charge created by the passage of the first. The first electron

[41] Type one superconductivity only occurs at temperatures close to absolute zero. For example aluminium is only a superconductor at temperatures below 1.14 K and lead at 7.19K (Tipler 1991, 1313)

moves through the lattice and distorts it by attracting the positively charged atoms towards it. This creates a trough of positive energy into which another electron is drawn, before the atoms spring back into their normal position. The distortion of the lattice means that electrons, which would usually repel each other, travel in pairs, called “Cooper pairs” (see figure 3.2). These Cooper pairs interlock to form a pattern spanning the whole superconductor. The interlocked electrons do not scatter, and move coherently as a unit (Auyang 1998, 180). Keeping things cool is important because it reduces the molecular motion - when the temperature is raised and the lattice vibrates more, the electron pairs and therefore the interlocked network of electrons are broken apart.

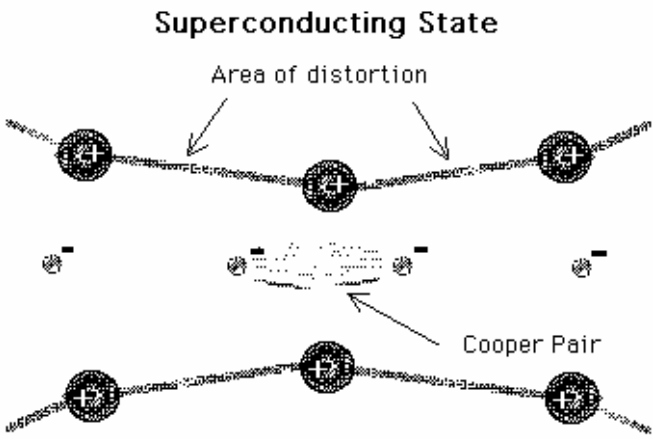


Figure 3.2 The distortion of the crystal lattice and the resulting formation of a cooper pairing of electrons (http://science.howstuffworks.com/framed.htm?parent=superconductivity.htm&url=http://www.ornl.gov/reports/m/ornlm3063r1/contents.html)

What we can see is that there is more to being a superconductor than having certain types of constituents, yet superconductivity can be explained in mereological terms. If you take certain metallic crystal lattices and cool them down, then they can do things that they cannot

do at higher temperatures. We can understand how these organised structures can do things at low temperatures that they cannot do at higher temperatures in terms of the behaviour of their individual constituents. Superconductivity provides another example of a complex phenomenon that has a different type of casual power to its parts, and yet can be explained in mereological terms.

3.2.3 Oxidative Phosphorylation [42]

Examples of explicable novelty can be found in biology as well as chemistry and physics. The formation of adenosine triphosphate (ATP), which is a biological fuel source, is a central part of metabolism of cells. ATP is produced within mitochondria via oxidative phosphorylation (the oxidation of sugars and fats). The key point we need to appreciate is that the process is dependent on the structure of the inner membrane of the mitochondria. The reactions would not be able to take place were it not for the membrane (see figure 3.3). Thus, a new type of chemical reaction is able to occur on the membrane. Nevertheless, this reaction is explicable.

The outer membrane of a mitochondrion is highly permeable whereas the inner membrane is highly impermeable, especially to ions. An electrochemical gradient is developed across the inner membrane by an electron transport system. Electron transport systems are composed of proteins and smaller compounds that have the ability to donate and accept electrons. Electrons move sequentially from high to low potential, each compound within the chain having a slightly greater affinity for electrons. As the membrane is relatively impermeable, the result of this electron chain is a proton gradient across the membrane. The synthesis of ATP begins with NADH and $FADH_2$, which temporarily store the electrons released in glycolysis and the citric acid cycle. NADH and $FADH_2$ transfer electrons to carrier proteins embedded in the inner membrane of the mitochondrion. Electrons from NADH and

[42] This example is taken from Bechtel and Richardson (1992). The biological details are taken from Knox et al (1994).

$FADH_2$ are removed and are passed through three protein complexes. Associated with this transfer of electrons is the translocation of protons from the matrix (the inside of the mitochondrion) to the outer side of the inner membrane. The proton concentration gradient created by proton pumping provides the electrochemical force to drive ATP synthesis (see figure 3.3).

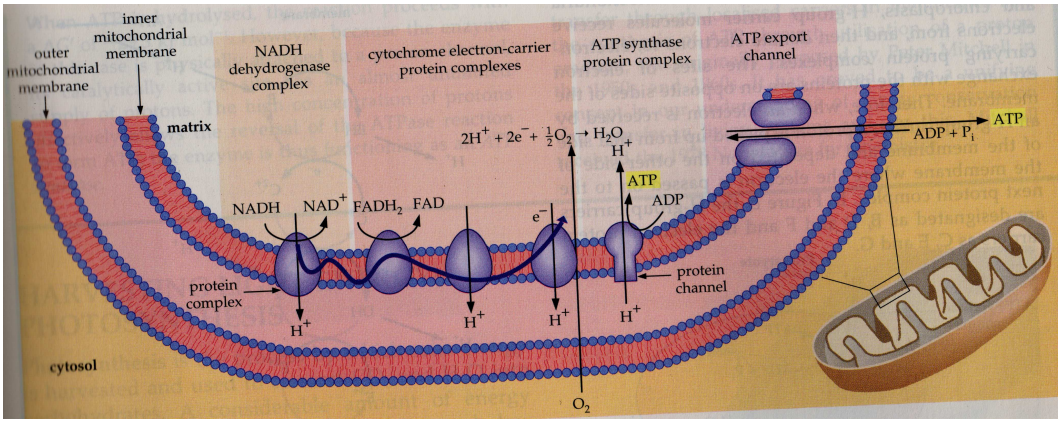


Figure 3.3 Synthesis of ATP on the inner membrane of the mitochondrion (Knox et al. 1994, 83).

As protons diffuse down the concentration gradient back across the mitochondrial membrane into the matrix, they pass through special protein channels. On the inner membrane of the mitochondria there is an enzyme that is attached to one of these channels. This enzyme, ATPase, normally catalyses the reaction:

$$ATP + H_20 \xrightarrow{ATPase} ADP + HPO_4 + H^+$$

However, the electrochemical gradient across the membrane creates a surplus of protons (H^+) on the outside of the inner membrane. The enzyme that catalyses the reaction is attached to a proton channel and its active site has an almost unlimited source of protons. This high concentration of protons pushes the equilibrium point of the reaction

towards the formation of ATP and enables the reversal of the reaction to form ATP.

Again we have an example of mereological novelty. The organised chemical system is able to do something, namely produce ATP via oxidative phosphorylation, which could not otherwise be done. Without the membrane structure, the process could not occur. The organisation of the membrane is crucial because the electrochemical gradient that is used to drive the endergonic reaction[43] is dependent on the membrane being impermeable to ions. Unlike some other biological reactions, oxidative phosphorylation will not occur if one just extracts the relevant enzymes and substrates and places them in a liquid milieu (Bechtel and Richardson 1992, 277). When arranged in an appropriate way, the lipids, proteins and enzymes are able to do something (namely reverse the reaction), which the same constituents without the arrangement cannot. As Bechtel and Richardson put it:

> The complexity [of the membrane] made possible sorts of phenomena which could not be generated by the components alone or when put together in a simple manner … This constitutes a sense in which the phenomena in question are emergent: they are different in kind from the phenomena that can be generated without the structured systems and can only be understood once we understand the structured systems (Bechtel and Richardson 1992, 278)

The fact that the chemical constituents of the membrane cannot perform oxidative phosphorylation unless they are arranged in the right way demonstrates that the arrangement of the chemicals into the membrane structure plays a crucial role in determining the causal powers of the membrane. The membrane can do things that its constituents in a different arrangement cannot. The membrane structure is mereologically novel.

Nevertheless, this mereological novelty is explicable. By arranging lipids, proteins and enzymes into a membrane with protein

[43] An endergonic reaction is one that absorbs energy.

channels, it is possible to produce a situation in which a chemical reaction (the synthesis of ATP from ADP) can occur. Nevertheless, this ability of the membrane to do things that its parts in a different arrangement cannot is not mysterious; we can see how the organised chemical structure that is the membrane is able to do things that its constituents cannot.

These examples of explicable mereological novelty suggest that we need to pull apart the two different themes that we found in examining the different conceptions of emergence. The claim that emergent wholes are inexplicable in mereological terms seems very questionable in light of current scientific work (McLaughlin 1992). However, the claim that there are mereologically novel phenomena that have causal powers that their parts lack has been corroborated by the above examples from various sciences. In the next section I want to consider whether it is possible to develop a metaphysics that accounts for this explicable mereological novelty.

3.3 Structural Emergence

We have seen that there is explicable mereological novelty. The challenge is to develop a metaphysics, and in particular an account of emergence, that can accommodate this. In this section I will argue that a structural account of emergence can meet our needs. According to this account complex phenomena have causal powers that their parts lack because they have causal powers that depend on the physical arrangements of the parts. I will also argue that a structural account of emergence does not have problems with providing for emergent causal powers, and is therefore immune to one of the more common criticisms of emergentist metaphysics.

3.3.1 The Account

Although the majority of emergentists understand emergence as both inexplicability and novelty, there have been some who have advocated a purely metaphysical understanding of emergence in terms of novelty (Auyang 1998; 2000; Bunge 1977; 1979; Wimsatt 1994; 1997). Going right back to the British emergentists, there has been the idea that the organisation of the parts is important in making the whole distinct from its parts (Auyang 1998; Bechtel and Richardson 1992; Broad 1925; Bunge 1979; Schroder 1998; Wimsatt 1997). The idea is that complex phenomena are distinct from their parts because the whole consists of the parts in a particular *arrangement*: 'emergent characters mostly belong to the *structural* aspect of systems and stem mainly from the organisation of their constituents' (Auyang 1998, 176). This “structural” account of emergence results in a mereological metaphysics, with metaphysical levels being defined in terms of parts and wholes (see figure 3.4). The whole consists of the parts in a particular arrangement and is ontologically distinct from the parts because it has causal powers in virtue of the arrangement of the parts.

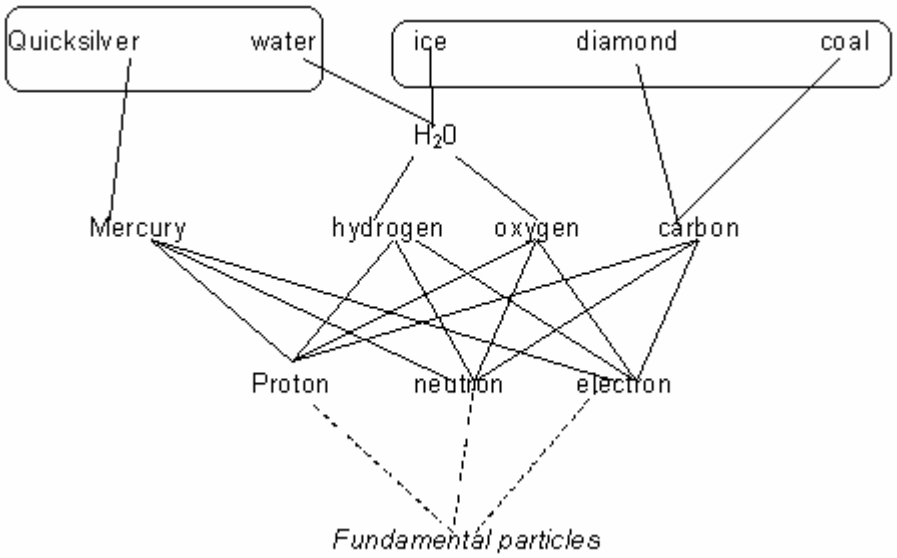


Figure 3.4: The configuration of the structuralist metaphysics. Parts are on one level, and organised

wholes on a higher level. The boxes on the top level are to indicate the commonality in causal powers between the things contained within the boxes. In this case the box on the left groups things that are liquids (at 1 atmosphere pressure and room temperature) whereas the one on the right groups those that are solids.

On a structural account of emergence not just any relationship among a set of parts is important. We can distinguish two different types of relations: causal and physical. Things can be organised functionally with causal relations, or structurally with physical relations.[44] Consider for example money. What makes something a unit of currency is nothing about the thing itself. Rather, what makes something money is the way in which it is used within an economy; if it is used by people as a store of value and is a medium of exchange, then it is money. If metals became particularly scarce, then there would be no reason why we could not switch and make our coins out of plastic instead. In general, it is the pattern of causes and effects that makes something money, so money is a functional phenomenon.

Compare this to an airfoil. Although an airfoil is defined as something that can provide lift, what enables something to provide lift is its shape. Lift is a force that is produced by turning a moving fluid (such as air). One way in which a fluid can be turned is by having something with a shape that distorts the flow of the fluid (see figure 3.5). What determines whether something can provide lift is its shape. What determines the shape of a complex object is not its *causal* relations to other things, but the physical relations of its constituents, such as the

[44] Note that it is far from clear that the causal relations that underlie functional phenomena are mereological. For example, the relevant causal relations for money are not those between the constituents of the physical entities that are the money. Rather, the relevant causal relations are between the physical entities that are the currency and the greater economic system. How to think metaphysically about functional phenomena is an important and complicated question. However it is *not* what I am examining here.

chemical bonds between the constituent atoms and the resulting spatial positioning of the atoms.

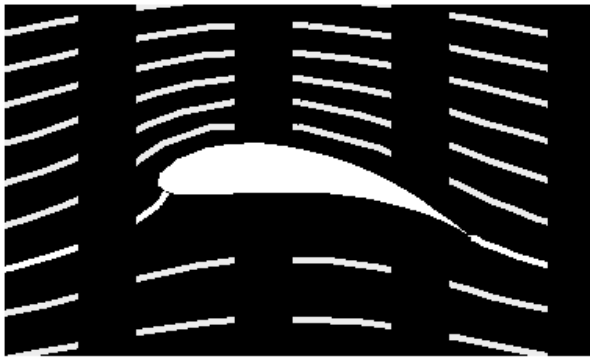


Figure 3.5: The differential flow of a fluid around an airfoil
http://www.grc.nasa.gov/WWW/K-12/airplane/factors.html

The key point is that on a structural account of emergence it is the *physical, not functional,* organisation that is important. In general, on a structural account of emergence, organised wholes have causal powers that their (un-organised) parts lack because the *physical* relations of the parts (partially) determine the causal powers of the whole. On a structural account of emergence, what matters is not causal, functional organisation, but physical, structural organisation.

We can see in the examples from section 3.2 that it is the *physical* arrangement of the parts in a complex phenomenon that underlies the difference between the causal powers of a complex phenomenon and its constituents. The reason why an organised collection of H_20 molecules has distinct properties and causal powers to those of its parts is because the collection has causal powers in virtue of the physical arrangement of the parts. The molecules in ice are bonded together with strong straight hydrogen bonds. In contrast, in liquid water, the open hydrogen bonded structure partially collapses and the molecules are closer to each other. This means that the molecules are closer to each other in liquid water and the liquid is denser than the

solid. So, ice floats on water. Simply by breaking or bending the strong straight hydrogen bonds[45] and moving the molecules closer to each other, we can change the causal powers of the substance. Importantly, the electrostatic bonds between the constituent molecules are not causal, functional relations. The molecules are not related functionally via cause and effect, rather they are bonded together by a physical (electrostatic) force. Collections of H_20 molecules have causal powers that depend on the nature of the physical relations between the molecules.

Physical organisation is also crucial to superconductivity. Superconductivity is crucially dependent on the organisation of the whole – you can only get a Cooper-pair between *two* electrons and superconductivity depends on interaction between *many* Cooper-pairs:

> The Cooper pairs interlock to form a pattern spanning the whole superconductor. Since the interlocking prevents the electrons from being individually scattered, the whole pattern of electrons does not decay but persists forever in moving coherently as a unit. The structure of the interlocked electronic pattern as whole, not the motion of individual electrons, is the crux of superconductivity (Auyang 1998, 180).

Superconductivity is a structural phenomenon: it is dependent on the *physical arrangement* of the constituents. The reason why temperature is important is because increasing the temperature destroys the structure that enables superconductivity. The relations between the constituents of a superconductor are physical relations. The crystal lattice and its constituent electrons are bonded together by electrostatic forces and it is in virtue of this overall physical organisation that the substance is able to conduct electricity with zero resistivity.

Finally, physical organisation plays a crucial role in oxidative phosphorylation. It is only because the inner membrane of

[45] This can be done either by increasing the temperature or by increasing the pressure (within limits).

mitochondria is impermeable to ions that the proton gradient that drives the reaction can be set up. Similarly, it is the organisation of the membrane, in particular the presence of the enzyme that catalyses the reaction *on* a transmembrane protein that transports protons, that provides the surplus of protons that is required to drive the reaction "backwards" to synthesise ATP. Again, it is physical relations, such as the sharing of electrons in covalent bonds, the spatial locations of ions and enzymes and the electrostatic forces between the constituents[46] that bind these elements together in the relevant structure that can do new things.

We can trace the differences in causal powers between a complex phenomenon and its constituents that were observed in section 3.2 to the physical organisation of the constituents. Basing mereological novelty in physical organisation also enables us to see how mereological novelty can be *explained*. The constituents of the novel phenomena that we examined above are able to form the structures that they do in virtue of their intrinsic natures. This is because the constituents are bound together with *physical* bonds. For example, it is the organisation of individual atoms that governs how they bond together and which therefore determines their causal powers. It is the polar nature of the H_20 molecule that is the basis of its ability to form hydrogen bonds. Metals are able to form crystal lattices because they have a relatively vacant outer shell. It is the organisation of the crystal lattice, and the ability of electrons to move freely throughout it, that underlies the superconductivity of metallic compounds. Finally, it is because the lipids that form membranes have hydrophilic heads and hydrophobic tails that they form a bilayer that is impenetrable (without the presence of channels or carrier proteins) to ions and large molecules.

The arrangement of the constituents of a structurally emergent phenomenon is determined by their physical nature. This means that it

[46] For example membranes such as the inner membrane of mitochondria consist of a lipid bilayer where the polar hydrophilic heads of the lipids are facing outwards (towards the surrounding water) and the non-polar hydrophobic tails of the lipids are facing inwards (away from the surrounding water).

is often possible to get a mereological explanation of structurally emergent phenomena. For example, it is ultimately possible to get at least the outline of an explanation of the bonding properties of metals and the nature of superconductivity in quantum terms. Similarly, as we saw above, it is possible get an explanation of why ice floats in water in terms of the nature of the H_20 molecule. In general, although the whole has causal powers that the parts lack, it is possible to account for the causal powers of the whole in terms of the nature of the parts. This is because the causal powers of the whole depend on the physical arrangement of the parts and the arrangement of the parts is determined by their intrinsic nature.

It is important to note that the mereological explanations that are possible on a structural account of emergence are not eliminative. Although it is possible to explain the whole in terms of its parts, higher-level theories are still required. For structurally emergent phenomena, not only does a similarity in constituents not guarantee a similarity in causal powers, a difference in constituents does not guarantee a qualitative difference in causal powers. Liquid water has much more in common with other liquids than it does with ice or steam. Despite their different chemical compositions, liquid water and ethanol share the causal powers that are characteristic of liquids. This is because they, like all liquids, have an organisation that is different from the lattice-like organisation of solids such as ice and iron, and the open random "structure" of gases such as gaseous water or carbon dioxide. This means that higher-level theories that take into account the similarity and differences in the organisation of the constituents are crucial. To treat any system of H_20 molecules as the same because it has the same constituents would be to miss important differences in causal powers that arise from the different arrangements of the constituents. In fact, if one merely focused on the similarities in the constituents, then the differences in causal powers that arise from the different possible organizations of the constituents would become inexplicable. A theory that was concerned only with the constituents would miss both similarities and differences between substances that are due to their

organisation. More than microphysics is needed to capture structurally emergent phenomena, because microphysics does not capture the emergent organisation.

On a structural account of emergence it is possible to get an explanatory "reduction" in the sense that mereological explanations are possible. However, this does not mean that we have an explanatory "reduction" in the sense that the higher-level theories become redundant and can be eliminated. This inability to eliminate higher-level theories is to be expected given that we cannot eliminate higher-level phenomena themselves. Bunge calls this "reduction without leveling":

> Explanation has not eliminated emergent properties … In particular, it has not eliminated the emergent laws characterising such systems: indeed such emergent laws are invariant relations among emergent properties, and emergence does not go away when explained. In other words, reduction does not imply leveling, it relates levels instead of denying that they exist (Bunge 1977, R79).

The fact that we can get mereological explanations of novel phenomena does not mean that we can bring the whole down to the same level as the parts. Nor does it mean that the higher-level theories of the whole become redundant. There is a difference between a mereological explanation and a reductive explanation that eliminates the higher-level theory (Bunge 1977, R79). In particular, we have seen that although it is possible to get mereological explanations of structurally emergent phenomena, it is not possible to eliminate the higher-level theories that concern these structurally emergent phenomena. A structural account of emergence explains the explicable mereological novelty that we found in the examples above.

Not only does a structural account of emergence capture the metaphysics of science, it can also deal with the emergentist's "trouble with causation" (Kim 1993b).

3.3.2 Multiple Realisability and Constitution Without Identity

In a series of papers, Kim (1992; 1993a; 1993b; 1993c; 1994) has argued that on the standard version of non-reductive materialism, there are no emergent causal powers and therefore there are no genuinely metaphysically emergent phenomena. In order to understand Kim's argument we first need to understand the view to which Kim is responding.

Particularly within philosophy of mind, it has become standard to base the *inexplicability*[47] of various phenomena in the fact that they are multiply realisable (Bickle 2002; Block 1997; Dretske 1990; Fodor 1974; 1997; Heil and Mele 1991; Horgan 1981; 1997; Le Pore and Loewer 1987; 1989; O'Connor, T. and Wong 2005; Putnam 1975). In general, a phenomenon is multiply realisable if it can be instantiated, "realised", in many different arrangements of microphysical entities. Many of the phenomena that the sciences talk about are diverse from the point of view of microphysics. For example, ice and diamonds are both solids even though they are made up different types of atoms. Similarly an airfoil ("wing") can be made out of aluminium, wood and canvas or even bone, muscle and feathers. Money is an even more striking example. A moment's reflection on all of the different types of currency shows that they need have no intrinsic physical properties in common. What matters is the way in which the currency is traded for goods and services, not its intrinsic physical nature. Phenomena such as money, solidity and airfoils are "multiply realisable" (MR) in that different arrangements of microphysical things can all be instances of the phenomenon.

It has been argued that the multiple realisability of various phenomena means that it is impossible to provide reductive explanations of them (Fodor 1974; Putnam 1975). The problem is that many different types of microphysical entities can share the same

[47] It is often far from clear whether such people also think that multiply realisable phenomena are somehow metaphysically distinct as well. The focus of the discussion is on the explanatory relevance of higher-level phenomena rather than their metaphysical uniqueness.

property (see figure 3.6). There seems to be no reason to think that everything that can act as a form of monetary exchange has anything in common from the point of view of physics. Similarly a bird's wing, airplane wing and sail do not seem to have anything in common from the point of view of microphysics. In general, it is claimed that MR phenomena are heterogeneous, disjunctive kinds from the point of view of physics. But, if MR kinds are heterogeneous and disjunctive from the point of view of physics, then it seems that it will not be possible to form a physics theory of them. If all the things that can be money have nothing in common from the point of view of physics then it is not possible to form physics laws about money. Disjunctive, ad hoc kinds do not provide counterfactually supporting generalizations and so do not support genuine laws (Fodor 1974; 1989; 1997). From the point of view of physics we are simply blind to the fact that a subset of all the physical things is money because from the point of view of physics money is not a kind.[48] It seems that it is not possible to identify multiply realisable *types* with types in microphysics.

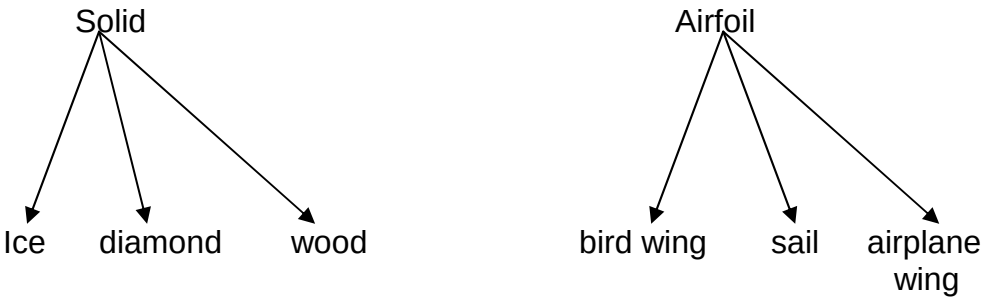


Figure 3.6: Multiple realisability is a relation that holds between a type and its various realisers. A type of phenomenon is multiply realizable if it can be realized in diverse ways.

[48] Kim (1993a) doubts that there can be higher-level counterfactually supporting generalisations if the realisation base is radically disjunctive. Although I think this may very well be right, as will become apparent below, I do not think that this is the primary problem.

Nevertheless, it is claimed that it is possible to get an identity of any particular multiply realisable special science event with a microphysics event. Although not all of the regularities in the world can be captured by physics, any particular event can be captured entirely using the science of physics (Fodor 1974). For example, although it is not possible to capture what all airfoils have in common from the point of view of microphysics, any particular wing, say the one belonging to the bird in the tree outside my window, can be described in purely microphysics terms. Similarly although liquidity cannot be identified with a particular molecular composition, the glass of water on my desk can be identified with a collection of H_20 molecules. It is claimed that any particular instance of a multiply realisable phenomenon is identifiable with its microphysics realiser.

Kim (1992; 1999) has argued that this identity of a token of a multiply realisable phenomenon with a microphysics phenomenon means that the token is always in competition for causal responsibility with its microphysics realiser. This follows directly from the fact that they are identical. Kim has summed up this problem in what he terms the “causal inheritance principle”: ‘if M is instantiated on a given occasion by being realized by P, then the causal powers of *this instance of M* are identical with (perhaps, a subset of) the causal powers of P' (Kim 1993b, 355). If every token of a multiply realisable phenomenon is identical with a microphysics entity then it is not clear what causal work there is left for the type to do.

The core difficulty is that multiple realisability is a relation that holds between a type of phenomenon and a particular instance of that phenomenon. However when we think about things in terms of causal powers it seems mistaken to say that a type has causal powers that are distinct from its realisers. This is simply because the causal powers that are characteristic of the type are equivalent to a subset of the causal powers of the individual realisers. There is nothing that liquids

can do that an individual, paradigmatic liquid such as water can't.[49] Similarly the causal powers that are characteristic of being an airfoil (such as providing lift) are in the particular wings themselves. When one focuses on causal powers one sees that the type belongs with the realisers. Where in the world is liquidity but in the liquids themselves?

This strongly suggests that MR phenomena do not have causal powers that are distinct from their microphysics realisers. This is a problem; if the type does no causal work of its own then, according to Alexander's dictum, it does not deserve a place in our metaphysics. If multiply realisable phenomena have no causal powers beyond their microphysics realisers, then in what sense are they distinct from their realisers? They certainly do not seem to be real if we accept Alexander's dictum as necessary for entry into out ontology.[50]

We can see that there is a tension between a standard argument for mereological inexplicability (which is based in multiple realisability) and a standard criterion of mereological novelty (possession of a different type of causal power). The core problem seems to be that, because any token of a multiply realisable phenomenon is identical to a physics entity, the token and its microphysics realisation base always seem to be in competition for causal relevance.

Adopting a structural account of emergence provides us with a form of non-reductive materialism that avoids these problems with causation. This is because it *denies* the identity of higher and lower-level tokens (and types). Recall that in a structuralist metaphysics, metaphysical levels are defined in mereological terms. At one level we

[49] Complications arise because our concepts many properties are more prototypical such that it is hard to give a set of necessary and sufficient conditions for membership in a certain class such as 'liquid'. But the inability to express in language what makes something a liquid does not in and of itself mean that there are no common causal powers that liquids share.

[50] Although it is fairly clear that possession of a different type of causal powers is *sufficient* for existence, it is far from clear that it is *necessary.* Thus, one standard reply to Kim has been to claim that it is not possession of causal *powers* that is necessary for existence, but rather causal *relevance* (Fodor 1989; 1997; Horgan 1997; Jackson and Pettit 1990; Le Pore and Loewer 1987; 1989; Yablo 1992). The worry with this move is that causal relevance seems to be an epistemic, not metaphysical factor. So, it is not clear that it will secure genuine metaphysical emergence. I do not want to adjudicate on this debate here. Suffice it to say that such a potentially problematic move is not necessary on a structural account of emergence.

have the parts (such as H_20 molecules) and at a higher level we have these parts in various arrangements (such as liquid water or ice). The key point is that the lower-level leaves out the organisation of the parts that is present in the higher emergent level. It is this arrangement which leads to the mereological novelty of the phenomena at the higher level. As such, the higher-level phenomena are not identical with their unstructured lower-level constituents. A structural phenomenon is not identical to its parts because simply specifying the parts does not specify their arrangement, which is crucial to determining the causal powers of the whole. A litre of (liquid) water is not identical to (approximately) $3x10^{25}$ H_20 molecules because $3x10^{25}$ H_20 molecules could be a solid, gas or pile of crystals. Similarly, we cannot identify a metallic superconductor with its constituent atoms because the same atoms at a different temperature are not a superconductor. In general, the organised whole cannot be identified with the parts at the lower level because the lower level leaves out the organisation that partially determines the causal powers of the whole.

This failure of identity across metaphysical levels that are defined in structural terms occurs not only for types but also for tokens. For example, the causal powers of the glass of water that is currently sitting on my desk are not identical to the causal powers of the constituent atoms because the same set of atoms could have different causal powers if they were arranged differently. Similarly a token superconductor cannot be identified with its constituent atoms because the same set of atoms in a different arrangement (say at a higher temperature) need not have the ability to conduct electricity with zero resistance. In general there is no identity across metaphysical levels for levels that have been defined in structural terms, because the lower level leaves out the organisation that (partially) determines the causal powers of the whole.

This failure of identities across metaphysical levels in a structuralist metaphysics means that higher-level emergent phenomena are not in competition for causal powers (or relevance) with their lower-level constituents. The key feature of the structuralist

metaphysics is that it provides constitution without identity.[51] Higher-level phenomena are constituted by the entities at lower levels, but they are in no way identical to the lower-level phenomena. Water is composed of H_20 molecules, but it is not identical to H_20. As water is not identical to its lower-level constituents, there is room for both water and its constituents to have their own unique causal powers. In fact, as we saw above, it is actually the case that water (i.e. a group of H_20 molecules in the appropriate arrangement) has causal powers that are distinct from those had by H_20 molecules (in a different arrangement).

In order to provide for genuinely emergent causal powers, we need to move away from thinking of our metaphysical levels in terms of multiple realisabilitiy and towards thinking about things in terms of mereological, part-whole relations.[52] If higher-level phenomena are constituted by lower-level entities, but not identical to them, then they will not be in competition for causal responsibility. A structural account of emergence provides constitution without identity because complex structural phenomena are not identical with their constituents.

Adopting a structural account of emergence, according to which the difference between wholes and their parts is the physical organisation of the parts, enables us to account for the explicable mereological novelty that is found throughout the sciences. Organised structures are novel and distinct from their constituents because they have causal powers in virtue of their physical arrangement. Because the nature of the constituents is a determinant of the structures that can be formed, it is often possible to get mereological explanations of structurally emergent phenomena. Finally, because the organised structure is not identical to its constituents, they are not in competition for causal

[51] This locution comes from Pereboom (2002, 504), however, he does not develop a structural account of emergence.

[52] To be fair to Kim I should point out that he recognises the possibility that his arguments only work against a nonreductive materialism that is based in the multiple realisability of higher-level phenomena: 'It may well be that there are versions of nonreductive physicalism to which my considerations do not apply, at least not directly; I believe though, that they are relevant to many of the more popular and influential versions of it' (Kim 1993b, 343)

relevance. Thus a structural account of emergence is able to account for how there can be emergent causal powers.

3.4 The Emergence of the Colour Visual Facts

Now that we have an understanding of structural emergence in general, we need to demonstrate that the states of affairs that are the colour visual facts are structurally emergent. What we want to show is that the facts that are represented in colour experience are *new* different facts from those which can be represented via other means.

As we saw in chapter two, the "clearly objective" facts that are represented in colour visual experiences are quite complex. Colour visual experiences represent, amongst other things, facts about the reflectance of surfaces, the nature of illumination and the way in which the nature of the light hitting a surface depends on the angle of illumination and the three-dimensional arrangement of light sources and object. They also represent information about the three-dimensional shape of objects and about the surfaces such as whether they are smooth or rough. Consider again the state of affairs that has been captured in the photo in figure 2.7. As we saw in chapter two, the state of affairs captured in the image is quite complex. The photo includes a wealth of information about the spectral, spatial and figural properties of the car, the cement, and the surrounding vegetation and buildings.

The important thing to appreciate about the complex states of affairs that are represented in colour visual experiences, such as that depicted in figure 2.7, is that the state of affairs consists of the constituents *and* their physical arrangements. The constituents of the state of affairs include the objects and the illuminating light, which are related by physical relations such as spatial relations. For example the state of affairs that is captured in figure 2.7 includes the spatial positions of the car, the concrete, the trees and the houses, as well as direction of the illumination. If these spatial positions, that is, if these

physical relations, were changed, then the state of affairs and the resulting image would be different. Similarly, if a cloud went over the sun, and the nature of the illumination changed, the light being reflected from the surfaces would change. For example we may no longer be able to see the reflection of the vegetation on the bonnet of the car or the shadow cast by the car on the concrete. In fact, if the clouds were particularly heavy we may lose all information about the relative positioning of the sun and the other objects. So, the state of affairs is complex with constituents that are related by physical relations.

More importantly, the causal powers of this complex state of affairs depend not only on the nature of the constituents but also on their physical relations. If the spatial relations between the car and the camera changed, or if the nature of the illumination changed, then a different pattern of photons would fall on the lens of the camera. The behaviour of light within the state of affairs and therefore the image that is captured by the camera depends not only on the nature of the constituent objects but also on the spatial (physical) relations. If you change the relations, then you change the behaviour of the light and therefore change the nature of the image that the state of affairs would cause.

The fact that the constituents in this arrangement have different causal powers than the constituents in a different arrangement indicates that the complex state of affairs that is the constituents and their arrangements is metaphysically distinct from its constituents. The fact that the source of this metaphysical difference between the constituents and the complex state of affairs is the physical relations of the constituents indicates that the state of affairs is structurally emergent. In other words, the complex states of affairs that are represented in colour visual experiences are more than their constituents; they consist of the constituents and their physical relations. That is, the colour visual facts are structurally emergent.

Now that we have a proper understanding of the metaphysics of the colour visual facts we will be able to turn to consider their

epistemology. In particular, is it plausible that the only way for us to represent these facts is via colour experience? More importantly, can these facts be physical even if we can only represent them via experience?

Chapter Four

The Knowledge Argument and Materialism

The task of this chapter is to use the metaphysical account developed in the last two chapters to develop an account of the epistemology of facts that we can know only via experience that is compatible with materialism. In the first section I will develop a general account of the epistemology. In the second and third sections I will consider whether this account is consistent with materialism. The second section will focus on whether a structuralist metaphysics is materialistic. In the third section I will consider whether the blind scientist's[53] ignorance prevents them from knowing that materialism is true and whether their ignorance means that their science is necessarily incomplete and that there are mysterious phenomena that they can never explain. We will see that it is possible to develop an account of colour vision on which there are physical facts that we can know only via colour vision. In the final section I will consider why philosophers have, for so long, been in the grip of the false intuition that facts that we can know only via experience cannot be physical.

[53] Note that I will talk about the blind scientist rather than Mary to make it clear that I am not claiming that the colour visual facts that I have developed an account of are the phenomenal colour facts. Recall that the aim of this thesis is to show that there can be physical facts that we can know only via experience. If this can be achieved, then materialism can be true even if the only way to know what an experience is like is to have the experience.

4.1 The Epistemological Account

In this section I will reconsider Mandik's general account of why there would be facts that we can know only via experience in light of the metaphysical account of the colour visual facts developed in the previous chapters. In particular, the issue to be examined is why it would be the case that there are some structurally emergent phenomena such as the colour visual facts that we can represent in only one way. In the first section I will consider some examples of the difference between our visual and non-visual representations of the colour visual facts. In the second section I will make a general suggestion concerning the inadequacies of non-visual representations of the colour visual facts. We will see that the difference between our visual and non-visual representations of the colour visual facts is that the latter are fragmented. The third section will provide a more detailed representational account of this difference, with the focus being on the representational resources of vision in comparison to other forms of mental representation. In the final section I will show how, given that the colour visual facts are structurally emergent, the fragmented non-visual representations of the colour visual facts are inadequate for knowledge of the facts themselves. We will see that, as the colour facts are structurally emergent, the problem is that to capture the complex fact in its emergent entirety requires representational resources which humans, with all of their epistemic restrictions, may only possess in one particular modality, namely vision.

Note that it is at this point that the account of the facts that we can know only via experience becomes particularly speculative and sketchy. This is an unfortunate consequence of the state of our knowledge concerning the representational resources of our brains. We simply do not know enough to be able to draw a clear distinction between the work done by visual and non-visual representations. However, this is not a deep problem. My aim is not to demonstrate that there must be facts that we can represent only in visual experience.

Rather, it is to see if it is possible to develop an account of such facts that is compatible with materialism. That is, I am not concerned with demonstrating that there *must* be facts that we can know only via experience. Rather, my goal is to show that materialism can be true even *if* there are such facts. To achieve this more modest goal it is only necessary to argue that the position I develop is plausible.

4.1.1 The Complexity of Visual Representations

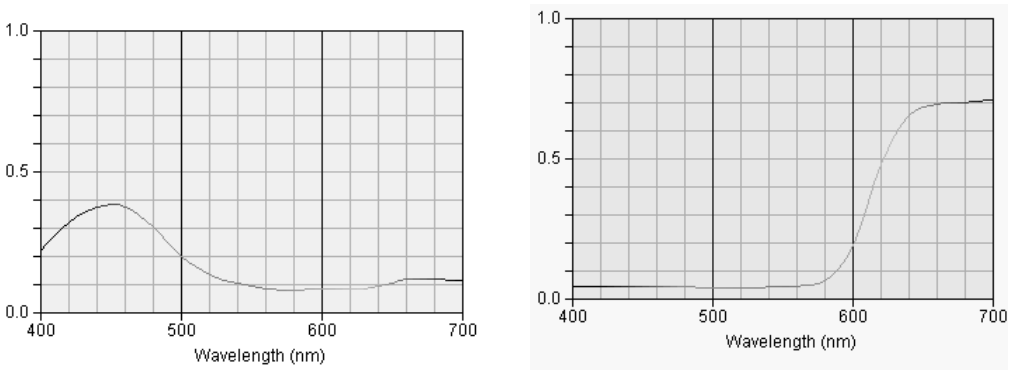
As I suggested at the end of chapter two, it is difficult to talk in general about the nature of the facts represented in colour vision. Instead, it is easier to talk about specific examples. I will examine some examples in order to see what the difference between our visual and non-visual representations of the colour visual facts might be.

Let's start with a simple example. Consider two rectangles placed side-by-side where the left one is blue and the right one is red. Now consider how we could go about representing non-visually, the sorts of things that according to the molecular account of vision our visual system is representing. We would need to start with some information about the percentage of photons reflected by the surfaces at different wavelengths. The spectral distribution curves included in figure 4.1 provide a sense of the quantity of information that could be obtained about the SSRs of the two surfaces.[54]

Next we would need to include some sort of comparison of the reflectances of the surfaces. Recall that it was argued that in order to achieve illuminant independent constancy the visual system takes into account the relation between the SSRs of adjacent regions (Gilchrist et al. 1999; Land, E.H. and McCann 1971; Lotto and Purves 2000; 2002; Rock 1975; Wallach 1948; Williams et al. 1998a; 1998b). We also need to compare the spectral reflectances of these two surfaces. One way to do this would be by calculating the difference in the percentage

[54] Note that I do not want to claim that the visual system represents as fine-grained information as is provided in the spectral distribution curves. How much of such information the visual system represents is an open empirical question.

of light reflected by each surface at various wavelengths. Given that the two coloured squares are surrounded by the white page, it may also be necessary to include a comparison of the SSRs of each the squares with the white page, as well as a comparison of all three together. Then there is also the surrounding black text. So, it may be necessary to do compare all four SSRs.

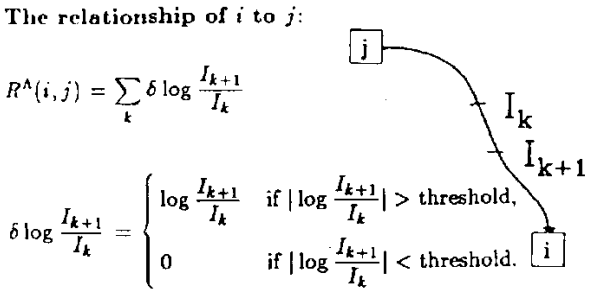


Blue Red

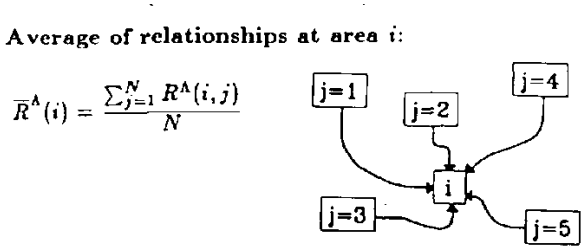
Figure 4.1 The spectral reflectances of a blue and red surface (Lindbloom 2005).

To do this non-visually would be quite laborious. Retinex equations provide an example of the complexity of such comparisons (Land, Edwin H. 1983; 1986).[55] These equations provide the means to exploit the fact that neighbouring surfaces are often affected by a change of illumination in the same way. As was pointed out in chapter two, although in bright light a black surface may reflect more light than a white surface in dim light, the ratio between the reflectances of the surfaces remains constant. Retinex equations essentially provide the means of performing such calculations for complex stimuli. First what needs to be calculated is the relationship between the flux at the target *i* in comparison to one of the surrounding regions *j*:

[55] Note that I do not want to imply that retinex equations necessarily capture how the visual system performs such calculations. They are simply included as an example of the type of things that the visual system computes.



Then what needs to be calculated is the average of the relationships between the target area *i* and each of the other areas in scene:



These comparisons are then carried out three times, once for each of the wavebands that each of the cones are sensitive to (Land, Edwin H. 1983; 1986).

However these equations are not sufficient to capture the information about the relations between the reflectances of surfaces that the visual system might calculate. One problem with retinex equations is that they assume that that any changes in illumination are gradual and therefore are eliminated by edge-based luminance ratios (Palmer 1999b). This assumption is often not borne out by the real world; sometimes shadows are sharp. The retinex theory doesn't to deal with these cases, although the visual system often does: colour constancy does not always fail when the visual system is presented with a surface under a shadow with a sharp edge (Palmer 1999b). Importantly for the current point, distinguishing between reflectance and illumination edges requires *additional* calculations, and therefore *increases* the complexity of the information that needs to be represented.

There are a number of heuristics that can be used to distinguish between reflectance and illumination edges. For example when a

shadow is cast across a surface, the two regions will tend to return the same shape spectral reflectance curve, the difference between the two regions simply being in the intensity of light that is reflected from each of them (see figure 4.2). In comparison, when an edge is due to a change in the reflectance of a surface, the reflectance spectra of the two surfaces will have a different overall shape (Palmer 1999b). So, in order to represent whether the differences between the regions are due to changes in surface spectral reflectances or changes in illumination, we also need to compare the overall shapes of the various luminance spectra of the various regions. Of course in this instance, we should come up with a result that indicates that there are no illumination edges in the visual scene. This information cannot be ignored because the visual system needs to take it into account in order to achieve illumination-independent constancy.

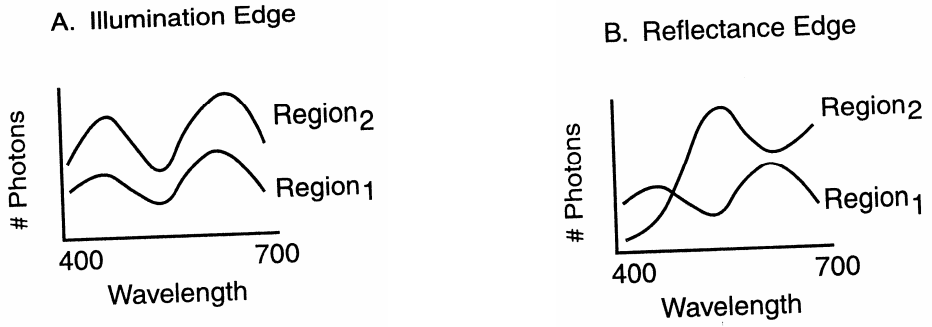


Figure 4.2 The difference in the shape of the SSRs for illumination versus reflectance edges (Palmer 1999b).

Finally, according to the molecular account of vision it is also necessary to describe some of the spatial and figural properties of the stimulus, such as that there are two coloured rectangles, each of which are 3.6cm wide and 3cm tall,[56] which are touching each other on a

[56] Note that it is also an open empirical question how precisely the visual system represents spatial properties.

white background, that the blue square is to the left of the red, and so on.

What we can see is that even for a simple stimulus (two coloured rectangles on a white background), there is a lot of information that the visual system could be representing. Moreover, in order to represent the same information non-visually, a great number of different representations would be required.

In comparison, the visual representation of the stimulus is quite simple:

When we move to a more realistic example, the complexity of the information is almost overwhelming. Consider the picture of the Jag in figure 2.7. Recall the complexity of the information that the visual system represents. Not only is there a great variety of different spectral reflectances, there is also a wealth of information about the spectral, spatial and figural relations between the various surfaces. Now think about how we could go about representing that information non-visually. Even the number of curves to capture the spectral reflectances of all of the different surfaces, such as the various shades of green of the leaves, is enormous. We also need some way to capture when the changes in reflectance are due to illumination changes and when they are due to changes in the nature of the surface. We can immediately “see” that the concrete is not painted two different colours, but rather that there is a shadow being cast on it by the car. The non-visual, scientific description would need to include not only information about the change in the nature of the light being reflected from the concrete, but also information about how the change is due to a change in illumination.

Information about the interreflections between the car and the surrounding trees and buildings also needs to be included. We can see that the bonnet of the car is a uniform material and that the change in the light leaving the bonnet of the car is due to it reflecting light that is being reflected from surrounding surfaces. It does not look like the car has trees and buildings painted on it, rather it looks like the top of the car is so well polished that it acts as a mirror. In order to capture this information non-visually, it is necessary to include information about the spatial relations of the car and the trees and house it is reflecting light from as well as the sun. We would also need to include information about the material properties of the various surfaces, such as that the top of the car is made of a material (polished, painted metal) that can act like a mirror.

Beyond spectral information we also need to capture all of the spatial and figural properties and relations. Consider how complicated the geometric description of the shape of the fog light on the bottom of the front of the car would have to be. What about capturing the different quantities of water on the cement? The list of information that is encoded in the image is enormous and so would be the quantity of non-visual representations that would be required to capture the information. What this image makes clear is the richness of the information that is represented in colour visual experiences and the difficulty of devising non-visual representations of these facts. But what precisely is the difference between visual and non-visual representations of the colour vision facts?

4.1.2 Fragmented Representations

The above examples of non-visual representations of colour visual facts suggest that such representations are disjointed or fragmented in comparison to visual representations of the colour visual facts. It seems plausible that in vision we can represent the complex relations between the spectral, spatial and figural properties of the world as a

whole. Non-visually, it is possible to represent all of these features, but it has to be done in fragments.

In talking about fragmented representations I have in mind the Gestalt notion that perception is not atomistic, but rather holistic (Hochberg 1974; Palmer 1999b). According to Gestalt theories of perception, whole perceptions have their own structure that cannot be reduced to that of their parts. Consider for example the octagons in the top box in figure 4.3. The octagons by themselves have perceptual properties such as colour, size and position. As a collection they have other properties, such as curvature, that are not possessed by any individual octagon (Palmer 1999b). Importantly for us, because the curvature is only a feature of the collection of octagons as a whole, a set of representations of the parts, such as those in the bottom portion of figure 4.3, will not capture the curvature of the whole. Even adding in the individual relations to indicate which octagon is next to which octagon seems insufficient to capture the curvature of the whole. To capture the curvature, we need to present all the octagons together in relation.

What I want to suggest is that the difference between the visual and non-visual representations of the colour visual facts is similar to the difference between the representation on the top and those on the bottom of figure 4.3. The collection of representations of each octagon on the bottom of figure 4.3 is fragmented; it represents each part, but not the parts in relation. In contrast, the representation of all of the octagons together on the top is integrated; it represents the parts in relation and therefore captures the curvature of the whole.

Consider science as an example of a domain where non-visual representation plays a crucial role.[57] It seems plausible that scientific

[57] Note that Paul Churchland (1995) has argued that in scientific thought we apply our perceptually acquired models to new domains. Although such a position makes scientific knowledge dependent on having experiences, it need not make it dependent on having a certain type of experience, such as colour experiences. The existence of congenitally blind scientists (such as Geerat Vermeij, a marine biologist) and even a colour-blind colour scientist (Knut Nordby), suggests that it is not necessary to have colour vision to learn science. Presumably, on Churchland's account, these scientists might be utilising their non-colour visual models in scientific thought.

representation is fragmented. Think about what happens when learning “the facts” of any science. When you walk into the lectures at the beginning of the year you are lead slowly through what we know about a phenomenon, investigating different aspects at different times. You don’t (and perhaps can’t) learn everything at once; rather you learn different bits at different times.

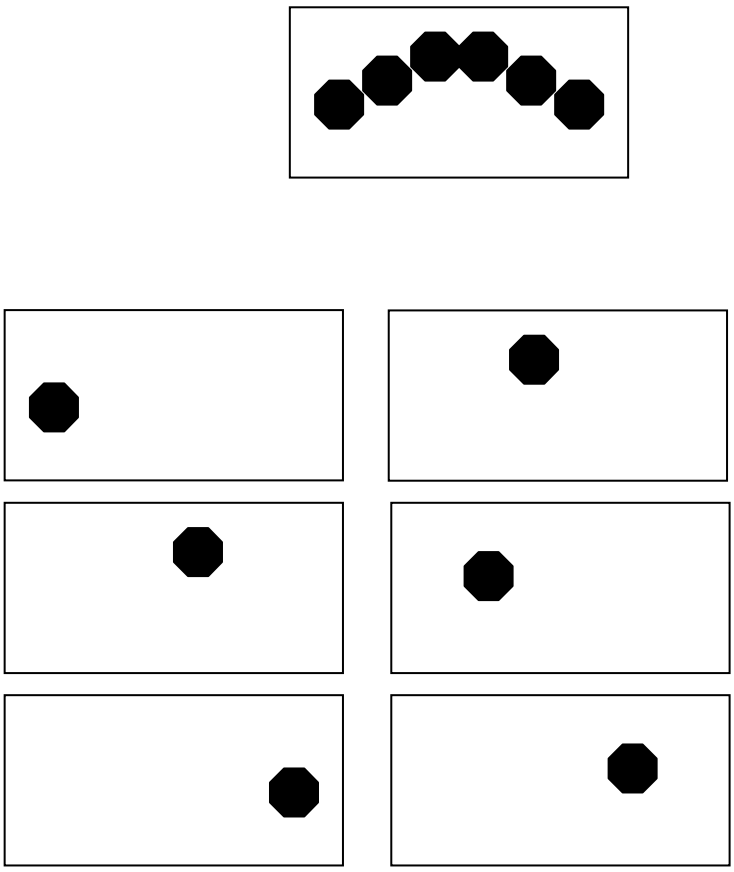


Figure 4.3 A perceptual gestalt: the octagons on the top have properties (such as curvature) which individual elements on the bottom lack.

For example, when learning about magnetic fields you may start by learning about their causal powers, then move on to investigating their mathematics, and finally you may learn about the virtual photons that underlie magnetic fields. Similarly, when learning about the autonomic nervous system you may start by learning about its

structural divisions, then work on the chemistry of each of the divisions, and finally learn about the function of each division as it relates, say, to controlling heart rate and digestive function. This breaking-up of a problem into smaller parts also occurred in our investigation of the colour visual facts. We were not required to represent a colour fact in its entirety, rather, we thought about various elements. First we looked at the effect of the spectral properties of the surround on the appearance of a surface. Then we investigated the role that spatial features of the stimuli play in determining colour appearance. Finally we investigated why it would be the case that the spectral and spatial properties of the surround determine the appearance of a surface. In doing this it was not necessary to represent the parts in relation. Rather, we represented the parts and their relations separately. This suggests that, at least when it comes to the colour visual facts, that our scientific representations in particular, and non-visual representations in general, are *fragmented*.

In contrast, if the above investigation of colour visual experiences is correct, then the experience does provide an integrated representation of the parts in relation. It is not the case that the experience represents the spatial properties separately from the figural properties, which are represented separately from the spectral properties. Rather colour vision represents the c between the spatial, figural and spectral properties as a whole.

The fragmented nature of non-visual representations of the colour visual facts is particularly evident in the examples considered above. We needed a series of representations of various aspects of the coloured squares to capture what the visual system can do in one representation. The picture of the jag makes the integrated nature of the visual representation even more apparent. The visual system can represent all of the relations between the spectral, spatial and figural features of the scene. In contrast, it seems that our non-visual representations of these features will be fragmented. Non-visually we can represent the different elements and some relations, but we cannot construct one representation of the parts in relation. The point is that

the amount of information within any visual experience is enormous. It certainly seems to outstrip the fragmented representations of the elements that can be obtained non-visually. So, it is plausible that our non-visual representations are inadequate because they are fragmented.

We need to be particularly careful when considering the differences between visual and non-visual representations. It seems that non-visually it is possible to represent not only the individual elements of the states of affairs that are the colour visual facts, such as surfaces and light sources. It also seems that we can non-visually represent the *relations* between the elements, such as their spatial relations, or the relations between the surface spectral reflectances. For example the retinex equations provide a means of representing the relative reflectances of various surfaces at different wavelengths. The suggestion is not that we cannot represent the parts and their relations. Rather, the suggestion is that we cannot represent the parts *in* relation. Consider how it is possible to represent not only the individual octagons in figure 4.3, but also their relations without capturing the curvature that only exists in the parts in relation. Similarly, it may be possible to non-visually represent the parts and their relations without capturing the parts in relation. For example we could plot the positioning of the octagons in a co-ordinate system or would could describe which octagon is next to which (if each frame is labelled a-f going left to right and top to bottom then the order is: a, d, b, c, f, e). We can have a non-visual representation of the spatial relation between a surface and illuminant, and a different non-visual representation of the surface spectral reflectance of the surface, and a third distinct non-visual representation of the nature of the light being reflected from the surface given the nature of the illumination. However, it seems that what we cannot represent, without having colour visual experiences, is the spatial relation between a surface and an illuminant, the surface reflectance of the surface and the nature of the light leaving the surface together. Why is it the case that our non-visual representations of the colour visual facts are fragmented?

4.1.3 Resources, Content and Fragmentation

We have seen that it is plausible that our non-visual representations of the colour visual facts are fragmented and that it is only in vision that we can represent the elements of the colour visual facts in relation. In this section I will speculate as to why non-visual representations must be fragmented. My general suggestion is that only the visual system has the resources required to represent the elements of the colour visual facts in relation. We will also examine the general sort of account that can be given on each type of content grounding relation.

Representation of facts as complex as the colour facts takes a lot of specialized representational resources. A more intricate represented object requires a more intricate representing system. It therefore seems plausible that the resources available to non-visual thought are inadequate to represent the colour visual facts. This claim is supported by the fact that over fifty percent of the cortex responds to visual input (Kaas and Krubitzer 1991; Kalat 1995; Palmer 1999b).[58] This suggests that vision's computational task is particularly complex. Why else would so much valuable neural architecture be devoted to it? My speculative hypothesis is that only the visual cortex has the specialized resources to represent not only the parts of the colour visual facts, but also the parts in relation of these emergent facts. The blind scientist's problem may simply be that, because they have been prevented from having colour visual experiences, they do not have the resources to build representations of the colour visual facts.

We can elaborate this general suggestion. Following Mandik (2001), we can say that, in general, the problem with non-visual representations is that they don't bear the appropriate content-

[58] Note that I am not necessarily claiming that all of these areas function purely to produce visual experiences. First of all, a lot of the areas included in this statistic do not respond only to visual information. Secondly, many of these areas, although involved in processing visual information, may not be responsible for the production of visual experiences. Nevertheless, if neural resources are the benchmark, then we are unquestionably visual creatures; no other sensory modality utilises such neural resources.

grounding relation to the colour visual facts. The precise details of why only colour experiences can bear the appropriate content grounding relation to the colour visual facts will vary depending on the type of content grounding relation that one's theory of content determination utilises.

Recall that a theory of content determination explains how a representing vehicle gets its content. As we saw in chapter one, although there are numerous theories of content determination, all the theories appeal to three different types of content grounding relations: causation, resemblance and convention. To give the flavour of the sort of explanations one could give, I will briefly sketch the type of explanation one could give for each *type* of grounding relation. The details of the account will depend on the specific theory of content determination one holds for visual experiences. For example the explanation given on a Fodorian indicator causal theory of content determination (Fodor 1987; 1990; 1994) will differ from that given on a causal teleological account of content (Dretske 1995; Millikan 1989; Papineau 2003; Sterelny 1990). Similarly the account provided on a structural resemblance theory of content determination (Cummins 1996; O'Brien and Opie 2004) will not be the same as that for a functional resemblance theory of content determination (Block 1986; Cummins 1989; Field 1978; Harman 1982; Loar 1982; McGinn 1982).[59] For our current purposes we can focus on the general types of content grounding relation rather than the specific theories of content determination.

First of all, given that we are concerned with knowledge, we are considering types of grounding relation that are plausible for mental representation. This means that we can immediately put convention to one side. No one has tried to develop a conventional theory of *mental* representation. Von Eckardt (1993) argues that this is because any theory of mental content determination that uses convention as the grounding relation is not naturalistic. A naturalistic ground of mental

[59] See Von Eckardt (1993, chapter six) for more on the different types of theories of content determination.

representation is one that does not implicate semantic or intentional properties (Von Eckardt 1993, 234). The problem is that it seems that a convention to use a particular type of representing vehicle (e.g. “tree”) to stand in for some represented object (e.g. “perennial plant with a self-supporting woody main stem”) implicates other representations. How can there be a convention without the *understanding* of the parties to the convention? If convention implicates understanding, then the convention depends on other mental representations that are involved in understanding. As such, convention is not a naturalistic grounding relation. In considering why it might be the case that only colour visual experiences can be about the colour visual facts, we can focus on causal and resemblance grounding relations.

In order to account for why only colour experience can be about the colour visual facts, theories of content determination that implicate causation can make use of Mandik’s (2001) notion that only the mechanism of colour vision can causally interact with the *whole fact*. The idea is that the colour visual facts are so complicated and “gerrymandered” that the only detection supporting causal interactions they can enter into are with colour visual system. Recall the torn piece of cardboard that spies used to use to recognise each other. The edge is so complicated that only the other half of the box will match it. The only mechanism for detecting one edge is the other edge. Similarly, the idea would be that the colour visual facts are so complicated, that the only representing producing mechanism that can causally interact with them is colour vision. For human beings, the only way to cause representations of the whole colour visual facts is to use your visual system. On theories of content determination that utilise causation, the resources that non-visual representation lack could be the causal mechanism of human colour vision.[60] On such an account, vision’s special resources are the causal mechanism that is the visual system.

[60] Note that I do not want to imply that theories of content determination which utilise causation *must* provide this sort of an account of Mary’s ignorance. Rather, this is intended to be one example of the sort of account that could be provided.

In order to explain why we can only represent the colour visual facts in colour vision on theories of content determination that make use of resemblance, one would need to provide some account of why colour visual experiences are the only mental representing vehicles that can resemble the colour vision facts. One option would be that there is something special about the nature of colour visual experiences themselves such that only they can resemble the colour visual facts. As an analogy, perhaps colour visual experiences are like curved blocks. Without using curved blocks there are simply some models you cannot construct, and so there are some things that you cannot represent. Restricting an individual to representations that do not include colour vision may be like restricting them to only using square blocks; there may be some representing vehicles that they cannot construct. Non-visual resources may be inadequate because they do not have access to the right *type* of representing vehicles.

Alternatively, it may not be that there is something special about colour visual experiences themselves; rather, the issue might simply be one of *quantity*. As I have just suggested, it is plausible that the visual cortex is able to devote considerable resources towards specialisation for a particular task. Perhaps the reason why only colour visual experiences can resemble the colour visual facts is because only colour vision has enough resources to be able to resemble the parts in relation. Perhaps it is like non-visual thought does not have access to enough curved blocks, or does not have enough space to build a model of sufficient complexity to resemble the colour facts. In general, for theories of content determination that utilise resemblance, colour visual experiences would be the only representing vehicles that could be about the colour visual facts because only colour vision provides the right type of resources or a sufficient quantity of resources to build model that can resemble the *whole fact*.

We can see that on the two types of content grounding relations that are feasible for mental representation, it is possible to account for why only colour visual experiences can be about the emergent colour visual facts.

Up until now the focus has been on the *representations* that are and are not about the colour visual facts. But, ultimately the issue is what we can *know* via experience. This focus on representation has occurred simply because it is far from clear what knowledge is (Steup 2006). What is generally accepted is that for an individual to have knowledge of a fact, they need to have a representation of ("belief about") the fact. As we saw when considering Mandik's (2001) reply, representation of a fact is necessary (although probably not sufficient) for knowledge of the fact. Although there is a gap that needs to be filled between representation and knowledge, it is not one that needs to be filled in order to make my argument go through. Given the arguments of the last chapters, it is plausible that there is a difference between our visual and non-visual knowledge of the colour facts because there is a difference between the visual and non-visual representations of these facts.

Now that we can see that there is a difference between the knowledge of the colour visual facts provided by vision and that which can be obtained non-visually, it is finally time to see how these different ways of knowing provide knowledge of different *facts*.

4.1.4 Knowing Structurally Emergent Phenomena

Given the metaphysical account of the colour visual facts developed in the preceding two chapters, it should be fairly obvious that fragmented non-visual knowledge of the parts and their relations is not sufficient for knowledge of the emergent states of affairs that are the parts in relation. The fact that the colour visual facts are structurally emergent and that only the visual system has the resources to be specialised to represent the constituents of the facts in relation means that there are facts that we can represent only via colour visual experiences.

We have just seen that it is plausible that, for human beings, only colour vision has the resources to represent the colour visual facts as a whole. Although it may be possible to represent fragments of the colour visual facts non-visually, it seems plausible that only colour vision can

represent the parts in relation. However, as we saw in the previous chapter, the colour vision facts are structurally emergent. The parts in relation are metaphysically distinct from the unrelated parts, or the parts in different relations. Given that the colour visual facts are structurally emergent, representations of the parts and their relations are not sufficient for a representation of the parts in relation. On a structural account of emergence, the parts in relation are mereologically novel. The colour visual facts are the parts in relation. To fail to represent the parts in relation is to fail to capture the fact that is the parts in relation.

Just as the set of “fragmented” representations of the individual octagons in figure 4.3 will not capture the curvature that is only in the octagons together, so too will a set of “fragmented” non-visual representations not capture the emergent state of affairs that are represented in colour vision. The non-visual representations of the parts and their relations are not a representation of the whole. Colour visual experiences enable us to represent new facts that we cannot represent non-visually.

We now have completed our account of the facts that we can know only via colour visual experiences. The facts that are represented in colour visual experiences are complex, emergent states of affairs concerning the co-dependence between the spectral, spatial and figural properties of the world. These facts can only be represented via colour visual experiences because only the visual system has the resources to represent such complex facts. Although it is possible to represent the constituents of these states of affairs without having colour experiences, we can not construct non-visual representing vehicles that bear the appropriate content grounding relation to any colour fact in its entirety. Because these colour visual facts are structurally emergent, the whole is metaphysically distinct from the parts. Therefore, the non-visual representations of the constituents of the colour visual facts are not representations of the colour visual facts as a whole. Thus, there are facts that we can represent only via colour visual experiences.

Now that we have developed our account of the facts that we can know only via experience, it is time to consider whether this account is consistent with materialism. Let's start with the structuralist metaphysics.

4.2 Structural Emergence and Materialism

Establishing that the colour visual facts are structurally emergent is a crucial part of my argument. However, it may be objected that it is far from clear that these emergent phenomena are physical. In this section I will argue that structurally emergent phenomena are physical.

It is notoriously difficult to provide a definition of the physical. In particular it is difficult to come up with a non-question-begging account of the physical that can act as an arbiter between different positions. This is because a conception of the physical is part of an overall metaphysics, and so cannot be divorced from a metaphysical view in order to arbitrate between competing accounts. This means that it is unrealistic to expect that I can, in a non-question-begging way, develop an account of the physical and then demonstrate that structurally emergent properties are physical. Instead, I will examine the metaphysics that arises out of a structural account of emergence (call this a "structuralist metaphysics") in more detail and argue that structuralism is distinct from both substance and property dualism. Note that I do not want to suggest that the metaphysics that is motivated by an examination of science can be completely captured in structural terms. For example, as I suggested in the last chapter, there may be dynamic forms of emergence as well as mereological ones. All I am claiming is that the view developed here is *part* of the story and that this part is consistent with materialism.

The simplest, most pure form of materialism is radically eliminative. It presents a flat metaphysics, where everything is nothing but the entities of microphysics. The chair on which I sit, the computer on which I am typing and my yawning office-mate, are all nothing but

the entities of micro-physics. On this view, even the more complex objects that physics deals with, such as atoms or molecules, would be nothing more than the entities of microphysics. For the extreme eliminativist, the only level in our metaphysics is the level of basic microphysics. On such an account, everything is physical in that everything is the entities of microphysics. For the radical eliminativist, the complexity we see in our world is apparent and the entities of microphysics are all that really exists.

Unlike the metaphysics of the radical eliminativist, the structuralist metaphysics is layered. Complex phenomena are emergent; they have different types of causal powers to those of their constituents. As such, structurally emergent phenomena are metaphysically distinct from their parts. These emergent phenomena can be placed on different levels on a part-whole basis.[61] Structured wholes have causal powers that their isolated constituents lack. So, the metaphysical levels in a structuralist metaphysics are (at least partially) mereological; entities at one level are the constituents of the emergent entities at the next level (see figure 4.3). There is more to a structuralist metaphysics than the entities of microphysics. Any form of materialism that is consistent with structuralism will be different to that presented by the radical eliminativist. In order to see how structuralism is a form of materialism, we need examine the metaphysical and epistemic relations between the various levels of the structuralist metaphysics.

[61] Note that I do not want to suggest that metaphysical levels can be defined completely in part-whole terms. One problem with a pure part/whole conception of levels is that it is not clear how to locate different complex systems that are not in a part-whole relation. For example, it is not clear what the level relationship between people and tables might be.

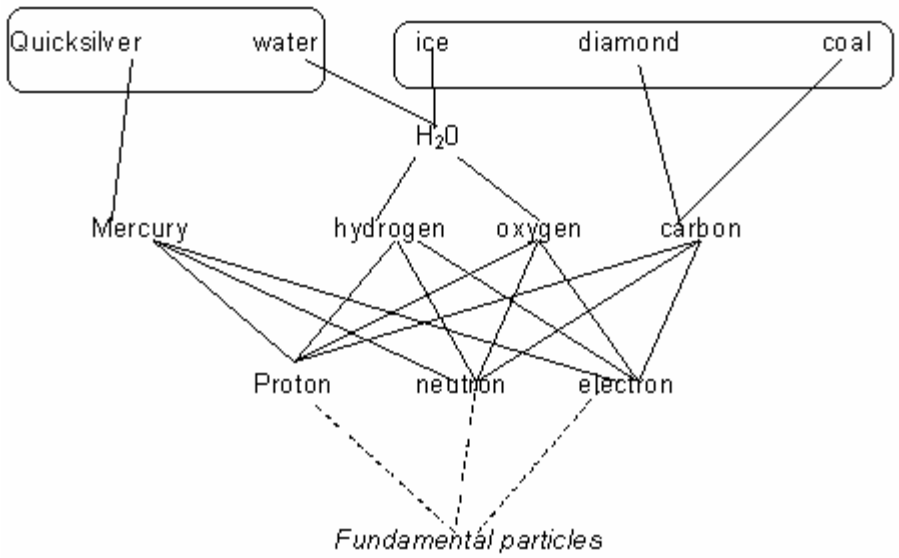


Figure 4.3: The configuration of (a small portion) of the structuralist metaphysics. The lower-level branches out as more structure is added. The dashed lines between the entities of microphysics and the next layer indicate that it is likely that a number of levels have been skipped.

We can see from figure 4.3 that underlying the diversity of a structuralist metaphysics there can be a fundamental unity. Ultimately there is one fundamental metaphysical level. More specifically, the structuralism we have been developing, which has been motivated by an examination of the sciences, is consistent with there being one fundamental metaphysical level that contains entities over which the science of microphysics quantifies. The structuralist can accept that everything is made up of the entities of microphysics: 'the various levels of organization are distinct, but they are all levels of organization of matter.' (Bunge 1977, R81-2). This is important, because it suggests that physics and the entities of microphysics are metaphysically fundamental in a way that is in some sense materialistic. Ultimately everything is made up of the entities of microphysics. If you took away the entities microphysics, then everything else would go as well. Thus structuralism is not a form of substance dualism.

The substance dualist claims that there are two fundamental types of substance in the world, the physical and the non-physical (usually the mental). When contrasted with substance dualism, materialism is usually taken to be a form of monism – ultimately there is only one type of substance: the physical. I suggest that conceiving of materialism as monism is too restrictive. To take a toy example, if the ultimate constituents of everything were protons, neutrons and electrons, it seems that there is a real sense in which materialism would still be true. What matters for the materialist is not so much that there is *one* fundamental thing, but that there is a fundamental set of things, all of which are within the reach of microphysics. Even the radical eliminativist could countenance the existence of more than one type of microphysics entity. It is only if there are ultimate constituents which are not within the reach of microphysics, that it seems sensible to classify the position as a form of substance dualism (see for example Chalmers 1996). What materialism requires is not so much monism as "(micro)physics foundationalism".[62] The basic entities, however many there are, all need to be within the reach of microphysics. Given this conception of substance dualism, we can see that structuralism is not a form of substance dualism, because ultimately everything is constituted by the entities of microphysics.

Is structuralism a form of property dualism? Like the materialist, the property dualist accepts that everything is constituted by the entities of microphysics. That is, the property dualist accepts physics foundationalism. However, the property dualist claims that these fundamental physics entities have non-physical properties. Defining "physical property" is no easier than defining "physical". However, I suggest that an examination of the history of science supports the view that obtaining a mereological explanation of a phenomenon is sufficient for us to accept it is physical. Consider life. At one point in time, vitalism was popular. Vitalism is the idea that the property of being alive is a non-physical property and that what turns inanimate matter

[62] Thanks to Jon Opie for suggesting this terminology.

into animate matter is the presence of a non-physical vital spirit (Robinson 2006). However, as biology progressed and we obtained a micro-explanation of life, vitalism fell out of favour. As we began to see how to build living organisms out of chemical and ultimately microphysics entities, we transitioned from thinking that the property of being alive was non-physical to accepting that the property of being alive was physical. In other words, obtaining a mereological explanation vanquished a property dualist account of life.

This reflection on the rise and fall of vitalism suggests that the property dualist claims that some properties are 'nonphysical in the sense that they cannot ever be reduced to or explained solely in terms of the concepts of the familiar physical sciences' (Churchland, P.M. 1988, 10 see also Auyang, 2000). What materialism requires is that there be a "perspicuous nexus" (Cottrell 1995) between the macro- and the micro-levels. We need to be able to see how emergent phenomena are a result of arranging the foundational things in the right way. I want to suggest that the property dualist claims that there are properties for which no such nexus exists and that the reason why we cannot know that such properties are physical is because they are not physical. The materialist claims that there are no properties for which there is a mystery. So, in order to distinguish materialism from property dualism it is necessary to consider the epistemic relations between levels. In particular we need to consider whether mereological explanations are possible.

As we saw in the previous chapter, it is possible to obtain a mereological explanation of structurally emergent phenomena. Recall that on a structural account of emergence, a complex phenomenon is distinct from its constituents because the nature (casual powers) of the emergent phenomena depends partially on the physical organisation of its parts. But, the organisation of the parts depends on their intrinsic physical nature. For example, many of the atypical properties of water arise from the hydrogen bonds that form between the molecules. The molecules form hydrogen bonds because they are highly polar. We can explain the properties of water in terms of its atomic constituents.

Although the whole has causal powers that the parts lack, it is possible to account for the causal powers of the whole in terms of the properties of the parts. This is because the causal powers of the whole depend on the arrangement of the parts and the arrangement of the parts is determined by their intrinsic nature. Mereological explanations are possible in a structuralist metaphysics.

The fact that it is possible to get mereological explanations of structurally emergent phenomena means that structuralism is distinct from property dualism. Although there are causal powers that do not belong to microphysical phenomena, we can get an explanation of how organised systems of microphysical phenomena are able to have causal powers that the isolated microphysical phenomena lack. This means that structurally emergent properties are not “mysteriously emergent” (Churchland, P.M. 1988), rather they are explicably emergent. There is no mystery, because we can see how the basic microphysics things, when arranged in the right way, can have emergent causal powers.

Structurally emergent phenomena are composed entirely of the entities of microphysics. This suggests that structuralism is distinct from substance dualism. The fact that we can, in general, see how to build the phenomena out of the entities of microphysics suggests that structuralism is distinct from property dualism. However, structurally emergent phenomena are not in any way identical to the entities of microphysics. They are metaphysically distinct and require their own theories. So, in what sense are they physical?

Pereboom suggests that the term “physical” has a broad and narrow sense: 'in the narrow sense, it picks out specifically anything over which theories in the science of physics quantifies; in the broad sense, it distinguishes anything wholly constituted from entities that are physical in the narrow sense' (Pereboom 2002, 526-7). We have just seen that structurally emergent phenomena will not be physical in Pereboom’s narrow sense; they are not (necessarily) entities that the science of physics quantifies over. But, structurally emergent phenomena are physical in the broad sense; they are wholly

constituted by the entities of physics. Although there is more to being water than being constructed of hydrogen and oxygen atoms, this extra bit is simply the arrangement of the parts. What is “added” to the entities of microphysics is “just” structure. According to the structuralist, everything is the entities of microphysics arranged in the appropriate way.

Importantly, the examples of structurally emergent phenomena provided in the previous chapter clearly demonstrate that science has already vindicated the broad conception of the physical. *Paradigmatically physical phenomena such as water, cells and ice are not physical in the narrow sense.* These phenomena are structurally emergent and therefore cannot be captured by microphysics. This means that if we adopt a narrow conception of the physical, water, ice and steam will not be physical. As far as I am concerned, this is a reductio of the narrow conception of the physical. Science shows us that not all paradigmatically physical things can be captured by physics. This strongly suggests that it is wrong to define the physical in terms of that which can be described by physics. Even if someone wanted to reserve the word “physical” only for narrowly physical phenomena, this would secure a merely verbal victory for the dualist. Most materialists would be happy if the mind had the same metaphysical standing as water. The fact that it is possible to develop a conception of the physical on which these things are not physical seems to be beside the point. The mind would be no more mysterious, spooky or non-physical than water. This would be a victory for materialism.

Structurally emergent phenomena are physical in the broad sense of being constituted entirely by the entities of microphysics. However, structurally emergent phenomena are not physical in the narrow sense of being within the domain of microphysics. The fact that water is only broadly physical strongly suggests that the broad conception of the physical is correct. Structuralism is distinct from substance dualism because it accepts physics foundationalism. It is distinct from property dualism because it admits of the possibility of mereological

explanations. So, structuralism is consistent with materialism. Everything in a structuralist metaphysics is constituted by the entities of microphysics and, because we can get mereological explanations we can see how structurally emergent phenomena arise out of the entities of microphysics. Accepting that there are facts that we can know only via experience does not support dualism because it requires an emergentist metaphysics.

4.3 Mary's Ignorance and Materialism

It is now time to face head on the worries about the facts that we can know only via experience and materialism. Is it really the case that materialism is consistent with there being facts that we can know only via experience? There seem to be two concerns: one about explanation; and one about our ability to know that materialism is true. One concern is that if we concede that there are facts that the blind scientist doesn't know, then aren't we conceding that there are facts that they cannot *explain*? That there are *mysterious* facts that are beyond the reach of their science? Wouldn't they then be faced with the prospect of property dualism? A second worry is that if we accept that there are facts that the blind scientist doesn't know, then how can they know that all the facts are physical facts? How can they know that materialism is true?

Both of these worries seem to apply to humans in general. There is a limit to the types of experiences that any human being can have. If there are facts that can be known only via experiences that we cannot have, then there are facts we cannot know. For example, if there are facts that can only be known via a bat's echolocatory experiences, then there are some facts that I cannot know. But if there are facts I cannot know, then how can I know that all of the facts are physical facts? Moreover, if there are facts that I cannot know, then how can I explain or understand these facts? How can I obtain the mereological explanations that are necessary to know that they are physical? If there

are facts that I cannot know, then aren't I doomed to be faced with mysteries?

I want to suggest that on the structuralist metaphysics developed in the previous chapter, these worries do not present any deep problems. This is because on this scientifically motivated metaphysics we get *explicable physical* mereological novelty. To take the worry about our ability to know that materialism is true first, the point to appreciate is that the broad conception of the physical is correct. The physical is that which is *constituted* by the entities over which the theories of microphysics quantify. To know that a phenomenon is physical one needs to know that it is *constituted* by the entities of microphysics. We have seen, however, that it is possible to have complete knowledge of the constituents of an emergent phenomenon, without knowing the phenomenon itself. This means that it is possible to know that a phenomenon is physical without knowing the phenomenon itself.

To return to the analogy of the curvature of the octagons, to know that the organised whole is physical what you need to know is that the constituents are physical, and that the emergence of the whole is due to the physical organisation of the parts. But, we can know that about the octagons without having to represent them as a whole, in relation. For example, we just have to know what each element is, and that the curvature comes from their spatial relations. Similarly, to know that the colour visual facts are physical, we do not need to be able to represent the colour visual facts themselves. Rather we need to be able to know the constituents of the colour visual facts, and more importantly, to know that ultimately, the colour visual facts are constituted entirely by the entities of microphysics. But, as we have seen, we can have complete knowledge of the constituents of the colour visual facts without knowing the facts themselves. We can know that the colour visual facts are nothing but the entities of microphysics arranged in the right way. In general, because it is possible to know the constituents of an emergent phenomenon without knowing the phenomenon itself, it is possible to know that a phenomenon is physical without knowing the

phenomenon itself. Knowledge of the truth of materialism does not require omniscience.

The second worry is that the blind scientist's science must be inadequate and that they will be doomed to be faced with mysteries because their ignorance will mean that there are facts that they cannot explain. If the above account is correct, then there are facts that the blind scientist doesn't know, and quite likely that we do not know either. If there are facts we don't know, then surely there are facts that we cannot explain. If there are facts that we can't explain, then how can we know that all of the facts are physical? Aren't we doomed to be faced with mysterious phenomena, which, as far as we can tell, might be non-physical? Given the means of distinguishing between property dualism and emergent materialism developed above, then if there are phenomena that we can't explain, it seems that we cannot rule out the possibility of property dualism.

On the other hand, it is hard to see how the blind scientist's science could be enhanced if they had their vision restored and are finally able to know the colour visual facts. This is why the intuition that Mary can gain complete scientific knowledge in her black and white room is so strong. It is hard to see how she could suddenly explain the colour visual facts any better. This suggests that there may be something mistaken with the above line of reasoning. As paradoxical as it first sounds, I want to suggest that on a structuralist metaphysics, it is possible to explain facts which you don't know.

First of all, with respect to the colour visual facts, if the explanation provided in chapter two is indicative, it seems that it is possible to explain them without having to represent them. By using other people as "measuring instruments" and sources of data, we can gain knowledge about the general types of things we are representing in colour vision (the spectral, spatial and figural features of the world). We can also know about the specific things people are representing in particular colour visual experiences (such as the relative spectral reflectances of two surfaces, or that there is a shadow which is making a surface appear two different shades, or that the positioning of a

shadow enables one to work out the relative spatial positioning of the objects). What we cannot do without having colour experiences is put the parts together in relation as a whole. However, this does not seem to be crucial to understanding the phenomenon that consists in the parts in relation. Rather what is required is comprehensive knowledge of the parts and their relations, which can be obtained without having colour experiences.

We can also see how it is possible to understand what a representational system is representing without having to have those representations by examining what we know about bats' echolocatory representations. In an attempt to understand what it is like to be a bat, Akins (1993) examines the nature of sound as a signal, the behaviour of bats, and bat neurophysiology.[63] Due to the incompleteness of our neurophysiological knowledge of the bat, the account is incomplete. Nevertheless, it goes a long way to providing insight into understanding what bats are representing in their echolocatory experiences, and there is no reason to think that a complete account is beyond our reach.

Bats use echolocation for space perception – if you cover the ears of a bat then it will fly about 'colliding at random with various objects' (Akins 1993, 128). Unlike human vision, a bat's echolocatory system makes use of sound waves, not light waves. The difference between the nature of the signals leads to important differences between spatial perception using light and that using sound waves. As sound waves involve the vibration of a medium such as air, sound has a high absorption rate. Compared to light, sonar signals cannot travel particularly far because the air rapidly absorbs them. This means that the bat is relatively myopic. So, simply on the basis of the signal being used we can determine that whatever bats are representing, it is relatively close to them. Another important difference between light and sound is that there is a relatively stable source of light (the sun) but

[63] Note that Akins ultimately argues that bats probably do not represent three-dimensional objective particulars in their echolocatory experiences. Rather, she suggests that echolocation is concerned with the "on-line" information processing that is required to navigate and catch prey. Note also that this is a brief summary of an admirably detailed paper.

there is no stable source of sound. This means that the bat has to produce its own sound waves.

The basic strategy of the bat is the parallel sequencing of signals and behaviour. They match the characteristics of the sound they produce to the task at hand; what the bat says determines what it hears (Akins 1993). Bats primarily use sonar when they are hunting. The principal purpose of the echolocatory system is to enable the bat to catch a moving object such as a moth, so echolocation enables the bat to represent small, fast-moving, fluttering objects like moths. They use two different types of signal: a constant frequency (CF) signal, which is ideally suited to measuring the relative velocity of the bat and its prey; and a frequency modulated (FM) signal, which is better suited for obtaining information about the distance to the target (Akins 1993, 134-5). Bats can alter the nature of the signal they produce in order to do different tasks. When searching for prey they emit a low frequency signal which travels further than a high frequency signal and is therefore well suited for obtaining information about distant objects. When they have detected prey and are approaching it they shift the signal so that it has more FM waves with a higher frequency, which are better suited to obtaining fine-grained spatial information and even textural detail about the prey. This suggests that, at different times during their hunting behaviour, bats are representing both the relative velocity and distance of their prey as well as fine-grained spatial information such as the texture of the target.

Electrophysiological studies of the bat's auditory cortex provide further insight into what bats are representing in echolocation. For example, there are neurons that become active when the returned echo rises above a certain threshold, that is, when the prey "comes into view". Akins suggests that the pattern of activity across these neurons could be used to provide information about the spatial location of the target. Information about the subtended angle of the target is obtained by monitoring the amplitude of the signals. By monitoring time differences between FM signals the bat is able to calculate the distance of its prey and by comparing CF signals the bat is able to obtain

information about the velocity of its prey (Akins 1993, 143-8). In general, it seems that the bat uses echolocation to obtain some very specific information about the spatial location and velocity of small moving targets.

The important thing about this example is that we are able to gain insight into what the bat is representing without having to have those representations ourselves. If it turns out that the features that the bat is representing are so complicated that we cannot represent them, then it may be the case that the bat is representing things we cannot.[64] What the example of the bat suggests is that in order to explain a representational system it is not necessary to have the representations of that system. Determining what the bat is representing does not require us to have the same representations as the bat. Similarly, as we saw in chapter two, working out the represented object of colour visual experiences does not require one to have colour visual experiences. Just as we can get an explanation of what the bat is representing without having the representations, so too can Mary get a full explanation of what it is that we are representing in colour visual experiences without thereby having to have the same representations. So, at least for representational systems, it seems that it is possible to explain a phenomenon (such as the represented objects of the system) without knowing the phenomenon.

Moving beyond explanations of representational systems, the key point to understanding how it is possible to explain something without knowing it is appreciating that mereological explanations abound in a structuralist metaphysics. First of all, it seems that it is possible to become aware of the existence of a phenomenon without having to represent the phenomenon itself. Recall that in the last chapter it was argued that the possession of causal powers that are distinct from those possessed by other things is a mark of existence. One reason why we would want to add something new to our scientific metaphysics is because we find that there are things that are being done which can't

[64] This is actually unlikely for bats. As Akins (1993) points out, bats are boring. They don't seem to represent particularly complex information.

be explained by those things which we already think exist. If we find a pattern of effects that cannot be explained by what we already know about, then we have good evidence that there is something new that we need to add to our worldview. But, it is not necessary to represent the phenomenon that is doing the new causal work in order to know about the causal powers themselves. It is possible to know that there is something that needs to be added to our metaphysics without actually knowing the phenomenon itself.

As we have seen above, it is also possible to have complete knowledge of the constituents and the relations of those constituents without knowing the emergent phenomenon that is the parts in relation. We can represent that there is something that is these parts in relation without representing the parts *in* relation. But, not representing the parts in relation is not representing the emergent phenomenon that is the parts in relation. Clearly it is possible to know what the phenomenon consists in without knowing the phenomenon itself. Finally, it is possible to explain the emergence of the phenomenon without knowing the phenomenon itself. To explain the emergence of a phenomenon we need to say how the parts in relation can do things that the parts not in that relation cannot. This requires understanding and representations of the parts and their relations, but not necessarily a representation of the parts in relation.

Consider the curvature of the octagons in figure 4.3. We can explain the curvature of octagons in terms of the pair-wise spatial positioning of the octagons. We don't need to represent all of the octagons in relation to be able to explain why they are curved. We can also see that knowledge of the parts in relation is not necessary to explain structurally emergent phenomena by considering the explanations of the emergent phenomena examined in the last chapter were like. For example, we saw how it is possible to explain why ice floats on water in terms of the strength of the hydrogen bond, which is in turn explained in terms of the nature of the hydrogen and oxygen atoms. This explanation does not require one to represent the H_20 molecules in the relations that they are in when they form ice. Rather it

just requires one to understand the general nature of the hydrogen bonds in ice, how the length of the bond varies with temperature and how density varies with the size of the intermolecular bond length. Similarly, we can have complete knowledge of the constituents of the colour visual facts and their arrangements without having colour experiences. As the novel causal powers that are characteristic of the colour visual facts arise because of the arrangement of the parts, this knowledge of the parts and their arrangements should enable us to explain the nature of the colour visual facts. In general, it seems to be possible to understand how a mereologically novel causal power can arise out of the organisation of a set of constituents without having to represent the emergent phenomenon that is the constituents in arrangement.

A little just-so story may help to make the account I am advocating clearer. Imagine that scientists become aware of a new pattern of effects. With some research they determine that this new causal work is not being done by anything which they already know about, so they hypothesise the existence of a new phenomenon, “x”, that is responsible for these hitherto undiscovered causal powers. Next they move on to trying to understand what x is. They manage to isolate three elements (“a”, “b” and “c”) which are necessary and jointly sufficient for the pattern of effects that are characteristic of x. However, they discover that the mere presence of a, b and c does not guarantee the presence of x (as detected by its characteristic effects). It is only in certain temperature and pressure conditions, that a, b and c form x. As they continue to manipulate the conditions under which a, b, and c, will form x they begin to understand that what is important is not only the presence of a, b and c, but also their arrangement. Further studies enable them to home in on the nature of the arrangements and how the organisation of the parts endows the whole with the causal powers that are characteristic of x. So, they conclude that x is a, b, and c in that particular arrangement. This seems to be an explanation of x. But, it does not seem that the scientists are required to represent a, b and c in relation. Rather, it seems that all that is required is an understanding of

the parts and their relations, and the unique causal work can be done by the parts when they are in that relation.

What these examples suggest is that facts that we lack the resources to represent are, like unobservable phenomena, explicable despite our "ignorance". Scientists are quite adept at dealing with unobservable phenomena. For example, our inability to perceive black holes or electrons has not prevented us from being able to explain them. I can gain information about the facts that a bat knows via echolocation even though I can't represent them, in the same way that I can gain information about electrons even though I can't represent them. Just as the unobservability of electrons or black holes does not put them beyond the reach of science or threaten the truth of materialism, so too does the unobservability of the facts that we can know only via experience not put them beyond the reach of science or threaten the truth of materialism. It seems that, at least on a structuralist metaphysics, it is not necessary to be able to know (represent) a phenomenon in order to be able to explain it.

We can now appreciate the problem with Jackson's conception of the physical. Although he does not attempt to define "physical", Jackson suggests that Mary has complete physical information in that she knows all of the information about the world provided by the physical, chemical and biological sciences and information (such as functional information) that can be inferred from this (1982, 127). Jackson's mistake is not so much to think that science plays a role in appreciating that a fact is physical. As we have seen, mereological explanations are important if we are to avoid property dualism. Presumably these explanations are scientific. Rather, Jackson's mistake is to think that it is only if a scientist can represent a fact that the fact is within the reach of science. Scientists can explain facts that they cannot actually represent. As long as they can represent the parts and the relations of a fact, they do not need to know a fact to be able to explain it or to be able to know that it is physical.

We can now see that accepting that there are facts that we can know only via experience does not mean that the facts cannot be

explained by those who cannot have the experience. This is because it is possible to know the constituents and relations of a phenomenon without knowing that constituents in relation that are the phenomenon. Not only does this knowledge of the constituents and their relations seem sufficient to explain the phenomenon. It also seems to be sufficient to know that the phenomenon is physical. This is because, on the broad conception of the physical to know that a phenomenon is physical we need to know that it is constituted by the entities of micro-physics. One need not know the phenomenon itself. Not only can there be physical facts that we can know only via experience, but it is also the case that we can know that materialism is true without knowing all the facts. Materialism is compatible with there being facts that we can know only via experience.

4.4 The Knowledge Intuition

I want to finish by speculating as to why the intuition that facts that we can know only via experience cannot be physical has gone relatively unchallenged for so long. Given how much attention has been paid to the knowledge argument, one might wonder why this type of reply is not more popular. It seems to me that the key overlooked step is the adoption of the proper, scientifically motivated metaphysics. In particular, the key thing about the structuralist metaphysics is that it is rich, yet the richness (mereological novelty) is explicable.

Materialism tends to be thought of as presenting a pretty sparse metaphysics (Robinson 2006; Stoljar 2005). Although many materialists claim to be non-reductive materialists, as we saw in chapter three, it is standard to base this claim in the multiple realisability of special science phenomena (Bickle 2002; Block 1997; Dretske 1990; Fodor 1974; 1997; Heil and Mele 1991; Horgan 1981; 1997; Le Pore and Loewer 1987; 1989; O'Connor, T. and Wong 2005; Putnam 1975). However, as we have seen, multiple realisability does not seem to provide a firm footing for genuinely emergent causal

powers. A number of participants in the debate over Mary also construe materialism in fairly reductive ways (Bigelow and Pargetter 1990; Chalmers 2004; Furash 1989; Jackson 1982; 1986; Lewis 1990; Nemirow 1990; Stoljar 2004; Stoljar and Nagasawa 2004). In other words, the potential richness of a materialistic metaphysics is not something that is generally recognised.

It is difficult, if not impossible, to accommodate the existence of facts that can we can represent only in experience in a sparse metaphysics. If the world is sparse or simple, then it seems that the representational resources required to capture all of the phenomena in the world would be relatively simple. In contrast, we have seen that a scientifically motivated metaphysics may actually be much richer than we initially thought. In particular, it seems that it may be so rich, and contain phenomena that are so complicated, that organisms such as us, with our epistemic restrictions, may simply not have the resources to represent many phenomena. It may very well be the case that the world is simply too rich for us to be able to capture it all.

We can see the importance of a rich metaphysics in making this type of reply to the knowledge argument work by considering the colour visual facts. As I have argued, it is plausible that although a colour blind scientist can know all of the parts and their relations, they cannot know the parts in relation. It is only if our metaphysics is rich enough to countenance a distinction between the parts and their relations and the parts in relation that it is possible that what is represented in experience is not what is represented scientifically. Similarly, if the states of affairs that are represented in a bat's echolocatory experiences are sufficiently complex that we are unable to represent the parts in relation, then even though we can get a description of the parts and their relations, we cannot represent the facts themselves. With a more sparse metaphysics, there would be no difference between what is known via representations of the parts and their relations and a representation of the parts in relations. With a metaphysics that is sparser than that which is motivated by a structural account of emergence, even if we admit that experiences provide a rich

representational medium, we are not forced to say that this representational medium provides knowledge of a new fact. It is only if the world is as rich as our experiences that our experiences may give us access to new facts.

However, not just any way of enriching the materialistic metaphysics is sufficient. As we have just seen, the details of the structuralist metaphysics are crucial for seeing that nothing is left out by science that does not directly capture emergent phenomena. First of all, the fact that the source of the metaphysical richness is "simply" the physical relations of the parts means that the position is not dualistic. A rich metaphysics is *not* obtained by adding more non-physical foundational entities or inexplicably emergent phenomena. More importantly, the fact that the source of the metaphysical richness is the physical relations of the parts enables us to make sense of how it is possible to explain emergent phenomena without actually representing them. For example, a colour blind scientist can represent the parts and their relations of the colour visual facts. This knowledge of the parts and their relations is sufficient to explain the phenomenon itself: all the colour visual facts consist in are the parts in relation. It is also sufficient to appreciate that such phenomena are physical. Yet, as we have seen, this knowledge of the parts and their relations is not sufficient for knowledge of the parts in relation. That is, it is not sufficient for knowledge of the facts themselves.

In other words, on a structuralist metaphysics we get not only metaphysical richness, but also the ability to explain the diversity without representing all the phenomena. It is possible to explain and understand the emergent complexity without representing it. This is because the basis of this metaphysical richness is an underlying simplicity, and it is possible to capture the simple elements and their relations without capturing the emergent complexity. On a structural account of emergence we get a metaphysics that is rich enough that we do not have the representational resources to know it all, but on which our inability to know everything is no bar to scientific explanation.

The key to appreciating that facts that we can know in only one way can nevertheless be physical is realising the explicable richness of the proper scientific metaphysics. Once we gain a more accurate understanding of the correct scientific metaphysics it becomes apparent that a scientist need not be able to represent a phenomenon in order to explain it. We were mislead into thinking that facts that we can represent in only one way cannot be physical because we failed to appreciate that our world is complex, yet explicable. Perhaps it is this ability to explain the phenomenon which underlies the intuition that we can have knowledge of it. We can now see that we can explain a fact without having to represent it. Explicability and knowledge of the fact itself can come apart in a structuralist metaphysics.

There can be physical facts which we can know only via experience. By adopting an emergent account of the facts that are represented in colour vision it is possible to show that the only way for a normal human being to represent those facts is via colour visual experiences. This is a crucial move in protecting materialism from the knowledge argument. Recall that the intuition that is at the heart of the knowledge argument is that the only way to know what an experience is like is to have the experience. This supposedly presents a problem for materialism because it seems that other physical facts can be known in many different ways. The difference between the way in which what an experience is like is known (only via experience) and the way in which other physical facts are known, suggests that they are themselves different.

I have shown is that this line of reasoning is not longer valid. By demonstrating that there can be *physical* facts that we can know only via experience, we can block the dualists move from an epistemic difference between what it is like and other physical facts to a metaphysical difference. This is because *there is no longer any epistemic difference.* There are physical facts that we can know in only one way. So, the (putative) fact that there is only one way to know the phenomenal quality of experience does not mean that it is not physical.

The move from epistemology to metaphysics that is made in the knowledge argument is no longer supported because there can be physical facts that we can know only via experience. The onus is not shifted firmly onto the dualist to show that materialism is incompatible with the knowledge argument because we can now see that materialism is true regardless of what happens to Mary.

Conclusion

The debate concerning the knowledge argument has reached a deadlock. Despite finding numerous different ways of responding to the knowledge argument, one can't help feeling that materialists have not made decisive progress. This is because the diversity of replies suggests that materialists have not found *the* response to the knowledge argument. I have argued that the problem is that materialists have failed to demonstrate that, regardless of what happens to Mary, materialism can still be true. We do not currently know what would happen to Mary upon release. Given this, the most comprehensive materialistic strategy is to demonstrate that, no matter what happens, materialism can still be true. Although materialists have found many ways to show that Mary's new knowledge may not be a problem, what they have not done was clearly demonstrate that materialism was compatible with Mary learning a new fact. This thesis has demonstrated that materialism is compatible with Mary learning a new fact. This has been achieved by demonstrating that facts that we can know only via experience can still be physical. The upshot is that even if the only way to know what an experience is like is to have the experience, this does not mean that what it is like is not physical.

We saw that, as Mandik (2001) suggested, one reason why there would be facts that we can know only via colour experience would be if

there were facts that can be represented only via colour experience. Moreover, there would be facts that can be represented only via colour experience if there is only one type of mental representing vehicle that can bear the appropriate content-determining relation to the colour visual facts. Perhaps surprisingly, we need to say very little about the nature of consciousness in order for this to be the case. In particular, all that is required is that representationalism be correct for colour experiences. As this is a position that is almost universally accepted, we can see that the knowledge argument does not tell us anything new about consciousness. Rather, what the argument suggests is that experiences provide a unique way of knowing the world. Accepting that there are facts that we can know only via experience therefore commits us to a number of claims about the facts that are known in colour experience.

In the second chapter we examined the nature of the facts that are represented in colour experience. By investigating the psychophysical data that has usually been taken to present a problem to the claim that colour experiences are representations, and by considering the ways in which the elements of a representational system might be dependent on each other for their content, we were able to see our way clear to a psychophysically plausible account of the represented object of colour visual experiences. In particular, I argued that colour vision is molecular in that each representing vehicle has its content only in virtue of its relations to the other representing vehicles. Thus it was suggested that colour vision provides a rich model of the visual world, in particular of the co-dependence between the spectral, spatial and figural properties of the world.

In the third chapter I argued that the colour visual facts are structurally emergent. I began by developing an account of mereological emergence in general. We saw that throughout the sciences there are examples of phenomena that are mereological novel, but explicable in mereological terms. We also saw that the source of this explicable mereological novelty is the physical organisation of the constituents. Thus, I argued that we should adopt a

structural account of (mereological) emergence according to which the whole is more than the sum of the parts because it consists in the physical organisation of the parts. It was then argued that, unlike the standard form of non-reductive materialism, which focuses on the multiple realisability of various phenomena, a structural account of emergence is able to provide for generally emergent causal powers. Finally, it was argued that the colour visual facts are structurally emergent.

In the final chapter I argued that the difference between our visual and non-visual knowledge of the colour visual facts is that the latter is fragmented. It is only in vision that we can represent not only the parts and their relations, but also the parts in relation. This is because it is only in colour vision that we have the resources required to capture facts that are as complex as the colour visual facts. Because the colour visual facts are structurally emergent, the parts in relation are metaphysically distinct from the parts and their relations. Thus, our non-visual knowledge of the parts and their relations is insufficient for knowledge of the emergent phenomena that are the parts in relation. So, there are facts that we can know only via experience.

With an account of the facts that we can know only via experience in place I then examined whether this account is compatible with materialism. We saw that structurally emergent phenomena are ultimately constituted by the entities of fundamental physics. As such, a structuralist metaphysics is not a form of substance dualism. We also saw that, because structurally emergent phenomena are mereologically explicable, it is possible to see how to build structurally emergent phenomena out of the entities of micro-physics. This means that a structuralist metaphysics is not a form of property dualism. Although structurally emergent phenomena are more than the entities of fundamental physics, they nevertheless seem to be physical. We also saw that the fact that there are facts of which the blind scientist would be ignorant need not prevent them from knowing that materialism is true. This is because they can know that the colour visual facts are constituted by physical entities without thereby knowing

the facts themselves. I also argued that on a structuralist metaphysics, it is possible to explain a phenomenon without knowing the phenomenon itself. The blind scientist's ignorance does not doom them to a science filled with mysterious, non-physics phenomena.

Finally I considered why we were mislead into thinking that facts that we can know only via experience cannot be physical. I suggested that ultimately the mistake we made was not to realise how rich a materialistic metaphysics can be. The world is complex and our minds are finite. We should not be surprised to find that there are things which are beyond our ability to know. However, the complexity is mereologically explicable and understanding the constituents out of which the complexity emerges is not beyond us. Despite our ignorance, the world need not be mysterious. We can maintain that our science can be explanatorily sufficient even though scientists are not omniscient.

It is possible for there to be physical facts that we can know only by having colour visual experiences. There are facts that we can know only via colour experience because it is only in colour vision that we can form representations that are sufficiently rich to capture the emergent structures that are the colour visual facts. These colour visual facts are nevertheless physical. This is because they are ultimately constituted by the entities of microphysics and we can see how to build them out of the entities of microphysics.

The upshot is that we have blocked the dualists move from our intuition that what it is like is known in a different way to other things to their conclusion that it is not physical. Even if the only way to know what an experience is like is to have the experience, this does not mean that what it is like is not physical, because we have developed an account on which there are physical facts which we can know only via experience. There is no longer an epistemic difference between what it is like and other physical facts. No matter what happens to Mary, materialism can be true. The dualist can no longer simply pump the intuition that Mary would be ignorant in order to justify their position.

Instead, the onus is now on the dualist to *justify* their intuition that Mary's ignorance is problematic for materialism.

The trouble for the dualist is that it is far from clear how they could do this. Although the emergentist metaphysics that has been developed in this thesis is novel, it is well supported by the empirical data. In order to undermine the current position and support their claim that Mary's ignorance is incompatible with materialism, the dualist now needs to *do the science* and show that the current position is not actually consistent with what science is telling us about the world. In other words, we can now move Mary out of the realm of armchair philosophy and into the realm of scientifically informed naturalistic philosophy. This, more than anything, is a victory for materialism.

More generally, what this examination of the knowledge argument suggests is that philosophers have set the bar too high when it comes to consciousness. When it comes to the knowledge argument, what is required to put consciousness on the same footing as other obviously physical phenomena is not a denial of our intuitions regarding consciousness, but a proper understanding of the metaphysics and epistemology of other physical phenomena. The highly reductive criteria that are often employed to demonstrate that consciousness is not physical would also lead to non-materialistic accounts of water, superconductivity and oxidative phosphorylation. Perhaps what is required to solve the hard-problem of consciousness is not a radical renovation of our intuitive conception of consciousness (à la Dennett 1991), but rather an elaboration of our conception of the physical to bring it in line with what science has already told up about the world. The problem in the hard problem of consciousness may not be consciousness, but our conception of the physical.

References

Agostini, T. and A. Galmonte (2002). "Perceptual Organisation Overcomes the Effects of Local Surround in Determining Simultaneous Lightness Contrast." *Psychological Science* **13**(1): pp.89-93.

Akins, K. (1993). What is it like to be Boring and Myopic. *Dennett and His Critics*. B. Dahlbom. Cambridge, Basil and Blackwell.

Akins, K. (1996a). "Lost the Plot? Reconstructing Dennett's Multiple Drafts Theory of Consciousness." *Mind and Language* **1**(1): pp.1-43.

Akins, K. (1996b). "Of Sensory Systems and the "Aboutness" of Mental States." *The Journal of Philosophy* **93**(7): pp.337-72.

Akins, K. (2001). More Than Mere Coloring: A dialog between philosophy and neuroscience on the nature of spectral vision. *Carving Our Destiny*. S. Fitzpatrick and J. T. Breyer. Washington, DC, Joseph Henry Press.

Akins, K. and M. Hahn (2000). The Peculiarity of Color. *Color Perception: Philosophical, Psychological, artistic, and Computational Perspectives*. S. Davis. Oxford, Oxford University Press.

Alexander, S. (1920). *Space, Time and Deity*. London, Macmillan.

Alter, T. (1998). "A Limited Defence of the Knowledge Argument." *Philosophical Studies* **90**: pp.35-56.

Arstila, V. (2003). "True Colors, False theories." *Australasian Journal of Philosophy* **81**(1): pp.41-50.

Atkins, P. W. and J. A. Beran (1989). *General Chemistry*. New York, Scientific American Books.

Auyang, S. Y. (1998). *Foundations of Complex-System Theories*. Cambridge, Cambridge University Press.

Auyang, S. Y. (1999). Synthetic Analysis of Complex Systems I & II. Presented at: How Scientists Deal with Complexity, University of

Sydney, June 15th 1999, http://www.usyd.edu.au/su/hps/newevents/complexity.html.
Auyang, S. Y. (2000). *Mind in Everyday Life and Cognitive Science*. Cambridge, MIT Press.
Averill, E. W. (1992). "The Relational Nature of Color." *The Philosophical Review* **101**(3): pp.551-88.
Barlow, H. B. (2001). Single Units and Sensation: A neuron doctrine for perceptual psychology? *Visual Perception: Essential readings*. S. Yantis. Philadelphia, Psychology press.
Bechtel, W. (2001). "The Compatibility of Complex Systems and Reduction: A case analysis of memory research." *Minds and Machines* **11**: pp.483-502.
Bechtel, W. and R. C. Richardson (1992). Emergent Phenomena and Complex Systems. *Emergence or Reduction? Essays on the Prospects of Nonreductive Physicalism*. A. Beckermann, H. Flohr and J. Kim. New York, Walter de Gruyter.
Bickle, J. (2002). Multiple Realizability. *The Stanford Encyclopaedia of Philosophy*. E. N. Zalta. **Fall 2002 Edition**.
Bigelow, J. and R. Pargetter (1990). "Acquaintance with Qualia." *Theoria* **56**(1-3): pp.129-147.
Block, N. (1986). "Advertisement for a Semantics for Psychology." *Midwest Studies In Philosophy* **10**: pp.615-78.
Block, N. (1990). "Inverted Earth." *Philosophical Perspectives* **4**(Action Theory and Philosophy of Mind): pp.53-79.
Block, N. (1996). "Mental Paint and Mental Latex." *Philosophical Issues* **7**(Perception): pp.19-49.
Block, N. (1997). "Anti-Reductionism Slaps Back." *Philosophical Perspectives* **11**: pp.107-33.
Brainard, D. H. (1998). "Color Constancy in the Nearly Natural Image 2: Achromatic Loci." *Journal of the Optical Society of America: A* **15**(307-25).
Brainard, D. H. and B. A. Wandell (1986). "Analysis of the Retinex Theory of Colour Vision." *Journal of the Optical Society of America: A* **3**(10): pp.1651-61.
Broad, C. D. (1925). *The Mind and its Place in Nature*. London, Kegan Paul, Trench,Trubner and Co.
Bunge, M. (1977). "'Levels and Reduction'." *American Journal of Physiology* **233**(3): pp.75-82.
Bunge, M. (1979). *Ontology II: A world of Systems*, D. Reidel Publishing.
Byrne, A. and D. R. Hilbert (2003a). "Color Realism and Color Science." *Behavioral and Brain Sciences* **26**: pp.3-64.
Byrne, A. and D. R. Hilbert (2003b). "Hardin, Tye and Color Physicalism." *The Journal of Philosophy*: pp.37-43.
Carney, T., M. Shalden, et al. (1987). "Parallel Processing of Motion and Colour information." *Nature* **328**: pp.647-9.
Cavanagh, P. and S. Anstis (1991). "The Contribution of Color to Motion in Normal and Color-Deficient Observers." *Vision Research* **31**(12): pp.2109-48.

Cavanagh, P., J. Boeglin, et al. (1985). "Perception of Motion in Equiluminous Kinematograms." *Perception* **14**: pp.151-62.
Cavanagh, P., C. W. Tyler, et al. (1984). "Perceived Velocity of Moving Chromatic Gratings." *Journal of the Optical Society of America: A* **1**(8): pp.893-9.
Chalmers, D. (1996). *The Conscious Mind: In Search of a Fundamental Theory*. Oxford., Oxford University Press.
Chalmers, D. (2004). Phenomenal Concepts and the Knowledge Argument. *There's Something About Mary: Essays on phenomenal consciousness and Frank Jackson's knowledge argument*. P. Ludlow, Y. Nagasawa and D. Stoljar. Cambridge, MA, MIT Press.
Chaplin, M. (2005). "Models for Water." Retrieved 23/05/2005, 2005, from http://www.lsbu.ac.uk/water/models.html.
Churchland, P. M. (1985). "Reduction, Qualia and the Direct Introspection of Brain States." *The Journal of Philosophy* **82**: pp.8-28.
Churchland, P. M. (1988). *Matter and consciousness : a contemporary introduction to the philosophy of mind*. Cambridge, Mass, MIT Press.
Churchland, P. M. (1989). *A Neurocomputational Perspective*. Cambridge, MIT Press.
Churchland, P. M. (1995). *The engine of reason, the seat of the soul : a philosophical journey into the brain*. Cambridge, Mass., MIT Press.
Churchland, P. M. (2005). "Chimerical Colors: Some Phenomenological Predictions from Cognitive Neuroscience." *Philosophical Psychology* **18**(5): pp.527-60.
Churchland, P. S. (1996). "The Hornswoggle Problem." *Journal of Consciousness Studies* **3**(5-6): pp.402-8.
Churchland, P. S. (2002). *Brain-wise : studies in neurophilosophy*. Cambridge, Mass., MIT Press.
Cohen, J. (2003). "Color: A functionalist proposal." *Philosophical Studies* **113**: pp.1-42.
Conee, E. (1994). "Phenomenal knowledge." *Australasian Journal of Philosophy* **72**(2): pp.136-148.
Cottrell, A. (1995). "Tertium datur? Reflections on Owen Flanagan's Consciousness Reconsidered." *Philosophical Psychology* **8**(1): pp.85-103.
Cropper, S. (2005). "The Detection of Motion in Chromatic Stimuli: first-order and second-order spatial structure." *Vision Research* **45**: pp.865-80.
Cummins, R. (1989). *Meaning and Mental Representation*, MIT Press.
Cummins, R. (1996). *Representations, Targets and Attitudes*. Cambridge, Bradford Books.
Cunningham, B. (2001). "The Reemergence of 'Emergence'." *Philosophy of Science* **68**(Proceedings): pp.S26-75.
Cunthasaksiri, P., H. Shinoda, et al. (2004). "Recognised Visual Space of Illumination: A new account of center-surround simultaneous

color contrast." *Color Research and Application* **39**(4): pp.255-60.
De Anna, G. (2002). "The Simple View of Colours and the Reference of Perceptual Terms." *Philosophy* **77**: pp.87-108.
Dedrick, D. (1996). "Can Color be Reduced to Anything?" *Philosophy of Science* **63**(proceedings): pp.S134-42.
Delahunt, P. B. and D. H. Brainard (2004). "Does Human Color Constancy Incorporate the Statistical Nature of Daylight?" *Journal of Vision* **4**: pp.57-81.
Dennett, D. C. (1991). *Consciousness Explained*. New York, Penguin Books.
Dretske, F. (1981). *Knowledge and the Flow of Information*. Cambridge, MA, MIT Press.
Dretske, F. (1986). Misrepresentation. *Belief*. R. Bodgan. Oxford, Oxford University Press.
Dretske, F. (1990). Does Meaning Matter. *Information, Semantics and Epistemology*. E. Villanueva. Oxford, Basil Blackwell.
Dretske, F. (1995). *Naturalizing the Mind*. Cambridge, MIT Press.
Edwards, R., D. Xiao, et al. (2003). "Color Sensitivity of Cells Responsive to Complex Stimuli in the Temporal Cortex." *Journal of Neurophysiology* **90**: pp.1245-56.
Field, H. (1978). "Mental Representation." *Erkenntnis* **13**: pp.9-61.
Fleming, R. W., A. Torralba, et al. (2004). "Specular Reflections and the Perception of Shape." *Journal of Vision* **4**(9): pp.798-820.
Fodor, J. (1974). "Special sciences, or the Disunity of Science as a Working Hypothesis." *Synthese* **28**: pp.97-115.
Fodor, J. (1987). *Psychosemantics: The problem of meaning in the philosophy of mind*. Cambridge, MA, MIT Press.
Fodor, J. (1989). "Making Mind Matter More." *Philosophical Topics* **42**(1): pp.59-79.
Fodor, J. (1990). *A Theory of Content and Other Essays*. Cambridge, MA, MIT Press.
Fodor, J. (1994). *The Elm and the Expert: Mentalese and its semantics*. Cambridge, MA, MIT Press.
Fodor, J. (1997). "Special Sciences: Still Autonomous After all These Years." *Philosophical Perspectives* **11**: pp.149-63.
Fodor, J. and E. Lepore (1992). *Holism: A shopper's guide*. Oxford, Blackwell.
Frege, G. (1984). Thoughts. *Collected Papers on Mathematics, Logic and Philosophy*. Oxford, Basil Blackwell.
Furash, G. (1989). "Frank Jackson's Knowledge Argument Against Materialism." *Dialogue* **32**(1): pp.1-5.
Gegenfurtner, K. R. and M. J. Hawken (1995). "Temporal and Chromatic Properties of Motion Mechanisms." *Vision Research* **35**(11): pp.1547-63.
Gegenfurtner, K. R. and D. C. Kiper (2003). "Color Vision." *Annual Review of Neuroscience* **26**: pp.181-206.
Gegenfurtner, K. R. and J. Rieger (2000). "Sensory and Cognitive Contributions of Color to the Recognition of Natural Scenes." *Current Biology* **10**: pp.805-8.

Gelb, A. (1938). Colour Constancy. *Source Book of Gestalt Psychology*. W. P. Ellis. London, Kegan Paul, Trench, Trubner.

Gilchrist, A. (1977). "Perceived Lightness Depends on Perceived Spatial Arrangement." *Science* **195**(4274): pp.185-7.

Gilchrist, A., C. Kossyfidis, et al. (1999). "An Anchoring Theory of Lightness Perception." *Psychological Review* **106**(4): pp.795-834.

Guillot, B. (2002). "A reappraisal of what we have learnt during three decades of computer simulations on water." *Journal of Molecular Liquids* **101**: pp.219-60.

Hanazawa, A., H. Komatsu, et al. (2000). "Neural selectivity for hue and saturation of color in the primary visual cortex of the monkey." *European Journal of Neuroscience* **12**: pp.1753-63.

Hardin, C. L. (1988). *Color for Philosophers*. Indianapolis, Hackett Publishing.

Hardin, C. L. (1990). Color and Illusion. *Mind and Cognition*. W. G. Lycan. Cambridge MA., Blackwell.

Hardin, C. L. (2003). "A Spectral Reflectance doth not a color make." *The Journal of Philosophy* **100**(4): pp.191-202.

Harman, G. (1982). "Conceptual Role Semantics." *Notre Dame Journal of Formal Logic* **23**: pp.242-56.

Heil, J. and A. Mele (1991). "Mental Causes." *American Philosophical Quaterly* **28**(1): pp.61-71.

Hempel, C. and P. Oppenheim (1948). "Studies in the Logic of Explanation." *Philosophy of Science* **15**(2): pp.135-75.

Heywood, C. A., R. W. Kentridge, et al. (1998). "Form and Motion From Colour in Cerebral Achromatopsia." *Experimental Brain Research* **123**: pp.145-53.

Hilbert, D. R. (1992). "What is Color Vision?" *Philosophical Studies* **68**: pp.351-70.

Hilbert, D. R. and M. K. Kalderon (2000). Color and the Inverted Spectrum. *Color Perception: Philosophical, Psychological, Artistic, and Computational Perspectives*. S. Davis. Oxford University Press, Oxford.

Hills, J. M. and D. H. Brainard (2005). "Do Common Mechanisms of Adaptation Mediate Color Discrimination and Appearance? Uniform Backgrounds." *Journal of the Optical Society of America: A* **22**(10): pp.2090-106.

Hochberg, J. (1974). Organization and the Gestalt Tradition. *Handbook of Perception. Volume One. Historical and Philosophical Roots of Perception*. E. C. Carterette and M. P. Friedman. New York, Academic Press.

Hoffman, D. D. (1998). *Visual Intelligence: How we create what we see*. New York, W. W. Norton & Co.

Horgan, T. (1981). "Token Physicalism, Supervenience and the Generality of Physics." *Synthese* **49**: pp.395-413.

Horgan, T. (1984). "Jackson on Phenomenal Information and Qualia." *Philosophical Quarterly* **34**: pp.147-152.

Horgan, T. (1997). "Kim on Mental Causation and Causal Exclusion." *Philosophical Perspectives* **11**: pp.165-84.

Horn, B. K. P. and M. J. Brooks (1989). Introduction. *Shape from Shading*. B. K. P. Horn and M. J. Brooks. Cambridge, MA, MIT Press.

Humphreys, P. (1997). "How Properties Emerge." *Philosophy of Science* **64**: pp.1-17.

Jackson, F. (1982). "Epiphenomenal Qualia." *Philosophical Quarterly* **32I**: pp.127-136.

Jackson, F. (1986). "What Mary Didn't Know." *Journal of Philosophy* **83**: pp.291-295.

Jackson, F. and P. Pettit (1990). "Program Explanation: A general perspective." *Analysis* **50**: pp.107-17.

Jimenez, J. R., M. Rubino, et al. (1997). "Chromatic Channels on Stereopsis with Random-dot Stereograms." *Vision Research* **37**(5): pp.591-6.

Kaas and Krubitzer (1991). The Organisation of the Extrastriate Visual Cortex. *Neuroanatomy of the Visual Pathways and their Development*. B. Dreher and S. R. Robinson. Boca Raton, CRC Press.

Kaiser, P. and R. Boynton (1996). *Human Color Vision*. Washington, Optical Society of America.

Kalat, J. W. (1995). *Biological Psychology*. Pacific grove, Brooks/Cole Publishing.

Kim, J. (1992). "Downward Causation" in Emergentism and Nonreductive Physicalism. *Emergence or Reduction? Essays on the Prospects of Nonreductive Physicalism*. A. Beckermann, H. Flohr and J. Kim. New York, Walter de Gruyter.

Kim, J. (1993a). Multiple Realization and the Metaphysics of Reduction. *Supervenience and Mind*. Cambridge, Cambridge University Press.

Kim, J. (1993b). The Nonreductivist's Troubles with Mental Causation. *Supervenience and Mind*. Cambridge, Cambridge University Press.

Kim, J. (1993c). Postscript on Mental Causation. *Supervenience and Mind*. Cambridge, Cambridge University Press.

Kim, J. (1994). The Myth of Nonreductive Materialism. *The Mind-Body Problem: A guide to the current debate*. R. Warner and T. Szubka. Oxford, Blackwell.

Kim, J. (1999). "Making Sense of Emergence." *Philosophical Studies* **95**: pp.3-36.

King, J. C. (2006). Structured Propositions. *The Stanford Encyclopaedia of Philosophy*. E. N. Zalta. URL = <http://plato.stanford.edu/archives/fall2006/entries/propositions-structured/>. **Fall 2006**.

Kingdom, F. A. (2003). "Color brings relief to human vision." *Nature Neuroscience* **6**(6): pp.641-4.

Kingdom, F. A. and D. R. Simmons (1996). "Stereoacuity and Colour Contrast." *Vision Research* **36**(9): pp.1311-9.

Kitcher, P. (1989). Explanatory Unification and the Causal Structure of the World. *Scientific Explanation*. P. Kitcher and W. Salmon. Minneapolis, University of Minnesota Press.

Kitts, D. B. (1977). *The Structure of Geology*. Dallas, SMU Press.
Knox, B., P. Ladiges, et al. (1994). *Biology*. Sydney, McGraw-Hill.
Kooi, F. and K. K. De Valois (1992). "The Role of Color in the Motion System." *Vision Research* **32**(4): pp.657-68.
Kraft, J. M. and D. H. Brainard (1999). "Mechanisms of Color Constancy Under Nearly Natural Viewing." *Proceedings of the National Academy of Science USA* **96**: pp.301-12.
Kraft, J. M., S. I. Maloney, et al. (2002). "Surface-illuminant Ambiguity and Color Constancy: Effects of scene complexity and depth cues." *Perception* **31**: pp.247-63.
Kraut, R. (1992). "The Objectivity of Color and the Color of Objectivity." *Philosophical Studies* **68**: pp.265-87.
Kuhn, T. S. (1962). *The Structure of Scientific Revolutions*. Chicago, University of Chicago Press.
Land, E. H. (1983). "Recent Advances in Retinex Theory and Some Implications for Cortical Computations: Color Vision and the Natural Image." *PNAS* **80**(16): pp.5163-5169.
Land, E. H. (1986). "Recent Advances in Retinex Theory." *Vision Research* **26**(1): pp.7-21.
Land, E. H. and J. J. McCann (1971). "Lightness and Retinex Theory." *Journal of the Optical Society of America* **61**(1): pp.1-11.
Laudan, R. (1978). "Philosophical Consequences of the Recent Revolution in Geology." *PSA* **2**: pp.227-39.
Le Pore, E. and B. Loewer (1987). "Mind Matters." *Journal of Philosophy* **84**(11): pp.630-42.
Le Pore, E. and B. Loewer (1989). "More on Making Mind Matter." *Philosophical Topics* **17**(1): pp.175-91.
Lehky, S. R. and T. J. Sejnowski (1988). "Network Model of Shape from Shading: Neural function arises from both receptive and projective fields." *Nature* **333**: pp.452-4.
Lehky, S. R. and T. J. Sejnowski (1990). "Neural Network Model of Visual Cortex for Determining Surface Curvature from Images of Shaded Surfaces." *Proceedings of the Royal Society of London B* **240**: pp.251-78.
Lennie, P. (1998). "Single Units and Visual Cortex Organisation." *Perception* **27**: pp.889-935.
Lewis, D. (1990). What experience Teaches us. *Mind and Cognition: A Reader*. W. G. Lycan. Cambridge, Basil Blackwell.
Li, X. and A. Gilchrist (1999). "Relative Area and Relative Luminance Combine to Anchor Surface Lightness Value." *Perception and Psychophysics* **61**(5): pp.771-85.
Lindbloom, B. (2005). Color Checker Calculator. http://www.brucelindbloom.com/. Accessed 15/1/2007.
Livingstone, M. and D. Hubel (1987). "Psychophysical Evidence for Separate Channels for the Perception of Form, Color, Movement, and Depth." *Journal of Neuroscience* **7**(11): pp.3416-68.
Livingstone, M. and D. Hubel (1988). "Segregation of Form, Color, Movement, and Depth: Anatomy, Physiology and Perception." *Science* **240**(4853): pp.740-9.

Loar, B. (1982). "Conceptual Role and Truth Conditions." *Notre Dame Journal of Formal Logic* **23**: pp.272-83.

Loar, B. (2004). Phenomenal States (Revised Version). *There's Something About Mary: Essays on Phenomenal Consciousness and Frank Jackson's Knowledge Argument*. P. Ludlow, Y. Nagasawa and D. Stoljar. Cambridge, MA, MIT Press.

Logothetis, N. K., P. H. Schriller, et al. (1990). "Perceptual Deficits and the Activity of the Color-Opponent and Broad-Band Pathways at Isoluminance." *Science* **247**(4939): pp.214-7.

Long, F. and D. Purves (2003). "Natural Scene Statistics as the Universal Basis of Color Context Effects." *Proceedings of the National Academy of Science USA* **100**(25): pp.15190-3.

Lotto, R. B. and D. Purves (2000). "An Empirical Explanation of Color Contrast." *Proceedings of the National Academy of Science USA* **97**(23): pp.12834-9.

Lotto, R. B. and D. Purves (2002). "The Empirical Basis of Color Perception." *Consciousness and Cognition* **11**: pp.609-29.

Lowe, E. J. (1995). Fact. *The Oxford Companion to Philosophy*. T. Honderich. Oxford, Oxford University Press.

Ludlow, P., Y. Nagasawa, et al., Eds. (2004). *There's Something About Mary: Essays on phenomenal consciousness and Frank Jackson's knowledge argument*. Cambridge, MA, Bradford Books.

Lycan, W. G. (2005). Representational Theories of Consciousness. *The Stanford Encyclopaedia of Philosophy*. E. N. Zalta. URL = <http://plato.stanford.edu/archives/spr2005/entries/consciousness-representational/>. **Spring 2005**.

Maloney, L. T. (1986). "Evaluation of Linear Models of Surface Spectral Reflectance with Small Numbers of Parameters." *Journal of the Optical Society of America: A* **3**(10): pp.1673-83.

Maloney, L. T. (2002). "Illuminant estimation as cue combination." *J Vis* **2**(6): pp.493-504.

Maloney, L. T. and J. A. Schirillo (2002). "Color Constancy, Lightness Constancy, and the Articulation Hypothesis." *Perception* **31**: pp.135-9.

Mandik, P. (2001). "Mental representation and the Subjectivity of Consciousness." *Philosophical Psychology* **14**(2): pp.179-202.

Marr, D. (1982). *Vision*. San Francisco, W.H. Freeman.

Marvin, U. B. (1973). *Continental Drift: The evolution of a concept*. Washington, Smithsonian Institution Press.

Matthen, M. (1988). "Biological Functions and Perceptual Content." *The Journal of Philosophy* **85**(1): pp.5-27.

Matthen, M. (1999). "The Disunity of Color." *The Philosophical Review* **108**(1): pp.47-84.

Matthen, M. (2001). "What Colors? Whose Colors?" *Consciousness and Cognition* **10**: pp.117-24.

Maund, B. (2006). Color. *The Stanford Encyclopaedia of Philosophy*. E. N. Zalta. URL = <http://plato.stanford.edu/archives/fall2006/entries/color/>. **Fall 2006**.

McGinn, C. (1982). The Structure of Content. *Thought and Content*. A. Woodfield. Oxford, Oxford University Press.
McGinn, C. (1989). "Can We Solve the Mind-Body Problem?" *Mind* **93**(391): pp.326-366.
McGinn, C. (1993). "Another Look at Color." *The Journal of Philosophy* **93**(11): pp.537-53.
McLaughlin, B. P. (1992). The Rise and Fall of British Emergentism. *Emergence or Reduction? Essays on the Prospects of Nonreductive Physicalism*. A. Beckermann, H. Flohr and J. Kim. Berlin, Walter de Gruyter.
Mill, J. S. (1879). *A System of Logic*. London, Longmans, Green and Co.
Millikan, R. (1989). "Biosemantics." *The Journal of Philosophy* **86**(6): pp.281-297.
Milner, A. D. and M. A. Goodale (1995). *The Visual Brain in Action*. Oxford, Oxford University Press.
Mollon, J. D. (1989). ""Tho' She Kneel'd in that Place Where the Grew": The uses and origins of primate colour vision." *Journal of Experimental Biology* **146**: pp.21-38.
Nagel, E. (1961). *The Structure of Science: Problems in the Logic of Scientific Explanation*. London, Routledge & Kegan Paul.
Nagel, T. (1974). "What is it like to be a Bat." *Philosophical Review* **83**: pp.435-450.
Nagel, T. (1986). *The View From Nowhere*. Oxford, Oxford University Press.
Nagel, T. (1998). "Conceiving the Impossible and the Mind-Body Problem." *Philosophy* **73**: pp.337-352.
Nemirow, L. (1990). Physicalism and the Cognitive Role of Acquaintance. *Mind and Cognition: A Reader*. W. G. Lycan. Cambridge, MA, Basil Blackwell.
Neumeyer, C., S. Dorr, et al. (2002). "Color Constancy in Goldfish and Man: Influence of surround size and lightness." *Perception* **31**: pp.171-87.
Nida-Rümelin, M. (2002). Qualia: The Knowledge Argument. *The Stanford Encyclopaedia of Philosophy*. E. N. Zalta. URL = <http://plato.stanford.edu/archives/fall2002/entries/qualia-knowledge/>. **Fall 2002**.
Nundy, S. and D. Purves (2002). "A Probabilistic Explanation of Brightness Scaling." *Proceedings of the National Academy of Science USA* **99**(22): pp.14482-7.
O'Brien, G. and J. Opie (2004). Notes Towards a Structuralist Theory of Mental Representation. *Representation in Mind: New Approaches to Mental Representation*. H. Clapin, P. Staines and P. Slezak, Elsevier.
O'Connor, T. (1994). "Emergent Properties." *American Philosophical Quarterly* **31**(2): pp.91-105.
O'Connor, T. and H. Y. Wong (2002). Emergent Properties. *The Stanford Encyclopaedia of Philosophy*. E. N. Zalta. **Winter 2002**.
O'Connor, T. and H. Y. Wong (2005). "The Metaphysics of Emergence." *Nous* **39**(4): pp.658-78.

Oxford-English-Dictionary (2006). Oxford English Dictionary, Oxford University Press.
Palmer, S. E. (1999a). "Color, Consciousness and the Isomorphism Constraint." *Behavioral and Brain Sciences* **22**: pp.923-989.
Palmer, S. E. (1999b). *Vision Science: Photons to phenomenology*. Cambridge, MA, MIT Press.
Papineau, D. (1993). "Physicalism, Consciousness and the Antipathetic Fallacy." *Australasian Journal of Philosophy* **71**(2): pp.169-182.
Papineau, D. (2003). "Is Representation Rife?" *Ratio* **16**(2): pp.107-23.
Papthomas, T. V., A. Gorea, et al. (1991). "Two Carriers for Motion Perception: Color and Luminance." *Vision Research* **31**(11): pp.1883-91.
Pereboom, D. (2002). "Robust Nonreductive Materialism." *The Journal of Philosophy* **99**(10): pp.499-531.
Pereboom, D. and H. Kornblith (1991). "The Metaphysics of Irreducibility." *Philosophical Studies* **63**: pp.125-45.
Pettit, P. (2004). Motion Blindness and the Knowledge Argument. *There's Something About Mary: Essays on Phenomenal Consciousness and Frank Jackson's Knowledge Argument*. P. Ludlow, Y. Nagasawa and D. Stoljar. Cambridge, MA, MIT Press.
Pokorny, J., S. K. Shevell, et al. (1991). Color Appearance and Color Constancy. *The Perception of Color*. P. Gouras. Ann Arbor, CRC Press. **6**.
Popper, K. R. (1963). *Conjectures and Refutations*. London, Routledge and Kegan Paul.
Popper, K. R. and J. C. Eccles (1977). *The Self and Its Brain*. New York, Springer.
Putnam, H. (1975). The Nature of Mental States. *Philosophical Papers II: Mind Language and Reality*. Cambridge, Cambridge University Press.
Ramachandran, V. S. (1987). "Interaction Between Colour and Motion in Human Vision." *Nature* **328**: pp.645-7.
Ramachandran, V. S. (1988). "Perception of Shape from Shading." *Nature* **331**: pp.163-6.
Ramachandran, V. S. and R. Gregory (1978). "Does Colour Provide an Input to Human Motion Perception?" *Nature* **275**: pp.55-6.
Ramsey, J. L. (1997). "Molecular Shape, Reduction, Explanation and Approximate Concepts." *Synthese* **111**(3): pp.233-51.
Revonsuo, A. (2001). "Putting Color Back Where it Belongs." *Consciousness and Cognition* **10**: pp.78-84.
Robilotto, R. and Q. Zaidi (2004). "Limits of Lightness Identification for Real Objects Under Natural Viewing Conditions." *Journal of Vision* **4**(9): pp.779-97.
Robinson, H. (2006). Dualism. *The Stanford Encyclopaedia of Philosophy*. E. N. Zalta. URL = <http://plato.stanford.edu/archives/fall2006/entries/dualism/>. **Fall 2006**.
Rock, I. (1975). *An Introduction to Perception*. New York, Macmillan.

Rosenthal, D. M. (2001). "Color, Mental Location and the Visual Field." *Consciousness and Cognition* **10**: pp.85-93.
Ross, P. (2001a). "Locating Color: Further Thoughts." *Consciousness and Cognition* **10**: pp.146-56.
Ross, P. (2001b). "The Location Problem for Color Subjectivism." *Consciousness and Cognition* **10**: pp.42-58.
Rossion, B. and G. Pourtois (2004). "Revisiting Snodgrass and Vanderwart's Object Pictorial Set: The role of surface detail in basic-level object recognition." *Perception* **22**: pp.217-36.
Russell, B. (1918). "The Philosophy of Logical Atomism." *Monist* **28**: pp.195-527.
Rutherford, M. D. and D. H. Brainard (2002). "Lightness Constancy: A direct test of the illumination-estimation hypothesis." *Psychological Science* **13**(2): pp.142-9.
Ruttiger, L., H. Mayser, et al. (2001). "The Color Constancy of the Red-Green Color Blind." *Color Research and Application* **26**: pp.S209-13.
Salmon, W. (1984). *Scientific Explanation and the Causal Structure of the World*. Princeton, Princeton University Press.
Scharff, L. V. and W. S. Geisler (1992). "Stereopsis at Isoluminance in the Absence of Chromatic Aberrations." *Journal of the Optical Society of America: A* **9**(6): pp.868-76.
Schiller, P. H. (1996). "On the Specificity of Neurons and Visual Areas." *Behavioral Brain Research* **76**: pp.21-35.
Schroder, J. (1998). "Emergence: Non-Deducibility or Downwards Causation?" *The Philosophical Quarterly* **48**(193): pp.433-52.
Searle, J. R. (1992). *The Rediscovery of the Mind*. Cambridge Massachusetts., MIT Press.
Shafer, S. A. and B. A. Maxwell (2000). Color as a Carrier of Physical Information. *Color Perception: Philosophical, psychological, artistic and computational perspectives*. S. Davis. New York, Oxford University Press.
Shepard, R. N. (1992). The Perceptual Organisation of Colors: An adaptation to regularities of the terrestrial world? *The Adapted Mind: Evolutionary psychology and the generation of culture*. J. H. Barkow, L. Cosmides and J. Tooby. Oxford, Oxford University Press.
Shoemaker, S. (1994). "Phenomenal Character." *Nous* **28**(1): pp.21-38.
Simmons, D. R. and F. A. Kingdom (1997). "On the Independence of Chromatic and Achromatic Stereopsis Mechanisms." *Vision Research* **37**(10): pp.1271-80.
Sinnott-Armstrong, W. and D. Sparrow (2002). "A Light Theory of Color." *Philosophical Studies* **110**: pp.267-84.
Smart, J. J. C. (1987). My Semantic Ascents and Descents. *Essays Metaphysical and Moral*. Oxford, Basil Blackwell.
Smithson, H. (2005). "Sensory, Computational and Cognitive Components of Human Colour Constancy." *Philosophical Transactions of the Royal Society: B* **360**: pp.1329-49.

Smithson, H. and Q. Zaidi (2004). "Colour Constancy in Context: Roles for local adaptation and levels of reference." *Journal of Vision* **4**(9): pp.693-710.

Sosa, E. (1995). State of Affairs. *The Cambridge Dictionary of Philosophy*. R. Audi. Cambridge, Cambridge University Press.

Stephan, A. (1992). The Historical Facets of Emergence. *Emergence or Reduction? Essays on the Prospects of Nonreductive Physicalism*. A. Beckermann, H. Flohr and J. Kim. Berlin, Walter de Gruyter.

Sterelny, K. (1990). *The Representational Theory of Mind*. Cambridge, MA, Basil and Blackwell.

Steup, M. (2006). The Analysis of Knowledge. *The Stanford Encyclopaedia of Philosophy*. URL = <http://plato.stanford.edu/archives/spr2006/entries/knowledge-analysis/>. **Spring 2001**

Stoljar, D. (2004). Two Conceptions of the Physical. *There's Something About Mary: Essays on Phenomenal Consciousness and Frank Jackson's Knowledge Argument*. P. Ludlow, Y. Nagasawa and D. Stoljar. Cambridge MA, MIT Press.

Stoljar, D. (2005). Physicalism. *The Stanford Encyclopaedia of Philosophy*. E. N. Zalta. URL = <http://plato.stanford.edu/archives/win2005/entries/physicalism/>. **Winter 2005**.

Stoljar, D. and Y. Nagasawa (2004). Introduction. *There's Something About Mary: Essays on Phenomenal Consciousness and Frank Jackson's Knowledge Argument*. P. Ludlow, Y. Nagasawa and D. Stoljar. Cambridge, MA, MIT Press.

Tanaka, J., D. Weiskopf, et al. (2001). "The Role of Color in High-Level Vision." *Trends in Cognitive Science* **5**(5): pp.211-15.

Thomas, N. J. T. (2001). "Color Realism: Towards a solution to the "hard problem"." *Consciousness and Cognition* **10**: pp.140-145.

Thompson, E. (1992). "Novel Colours." *Philosophical Studies* **68**: pp.321-349.

Thompson, E. (1995). "Colour Vision, Evolution, and Perceptual Content." *Synthese* **104**(1): pp.1-32.

Thompson, E. (2000). Comparative Color Vision: Quality space and visual ecology. *Color Perception: Philosophical, psychological, artistic, and computational perspectives*. S. Davis. Oxford, Oxford University Press.

Tipler, P. A. (1991). *Physics for Scientists and Engineers*. New York, Worth Publishers.

Turvey, M. T., V. Whitmyer, et al. (2001). "Explaining Metamers: Right Degrees of Freedom, not Subjectivism." *Consciousness and Cognition* **10**: pp.105-16.

Tye, M. (1995). *Ten Problems of Consciousness: A representational theory of the phenomenal mind*. Cambridge, MA, MIT Press.

Tye, M. (2000). *Consciousness, Color and Content*. Cambridge, MA, MIT Press.

Tye, M. (2004). Knowing What It Is Like: The Ability Hypothesis and the Knowledge Argument. *There's Something About Mary: Essays*

on Phenomenal Consciousness and Frank Jackson's Knowledge Argument. P. Ludlow, Y. Nagasawa and D. Stoljar. Cambridge, MA, MIT Press.

Van Brakel, J. (1997). "Chemistry as the Science of the Transformation of Substances." *Synthese* **111**(3): pp.253-82.

Van Gulick, R. (1985). "Physicalism and the subjectivity of the mental." *Philosophical Topics* **13**(3): pp.51-70.

Van Gulick, R. (2004). So Many Ways of Saying No to Mary. *There's Something About Mary: Essays on Phenomenal Consciousness and Frank Jackson's Knowledge Argument*. P. Ludlow, Y. Nagasawa and D. Stoljar. Cambridge, MA, MIT Press.

Von Eckardt, B. (1993). *What is Cognitive Science?* Cambridge, MA, MIT Press.

Wachtler, T., T. J. Sejnowski, et al. (2003). "Representation of Color Stimuli in Awake Macaque Primary Visual Cortex." *Neuron* **37**: pp.681-91.

Wallach, H. (1948). "Brightness Constancy and the Nature of the Achromatic Colours." *Journal of Experimental Psychology* **38**: pp.310-24.

Wandell, B. (1995). *Foundations of vision*. Sunderland, Sinauer Associates.

Watanabe, T. and P. Cavanagh (1992). "Depth Capture and Transparency of Regions Bounded by Illusory and Chromatic Contours." *Vision Research* **32**(3): pp.527-32.

Webster, W. R. (2002). "Wavelength Theory of Color Strikes Back: The return of the physical." *Synthese* **132**: pp.303-334.

Whewell, W. (1857). *History of the Inductive Sciences: From the earliest to the present time*. London, John W. Parker.

Williams, S. M., A. McCoy, et al. (1998a). "An Empirical Explanation of Brightness." *Proceedings of the National Academy of Science USA* **95**: pp.13301-6.

Williams, S. M., A. McCoy, et al. (1998b). "The Influence of Depicted Illumination on Brightness." *Proceedings of the National Academy of Science USA* **95**: pp.13296-13300.

Wimsatt, W. C. (1994). "The Ontology of Complex Systems: Levels of Organisation, Perspectives, and Causal Thickets." *Canadian Journal of Philosophy* **Supplementary** (Volume 20).

Wimsatt, W. C. (1997). "Aggregativity: Reductive Heuristics for Finding Emergence." *Philosophy of Science* **64**(Proceedings): pp.S372-84.

Wimsatt, W. C. (2000). "Emergence as Non-Aggregativity and the Biases of Reductionism." *Foundations of Science* **5**: pp.269-97.

Woodward, J. (2003). Scientific Explanation. *The Stanford Encyclopaedia of Philosophy*. E. N. Zalta. URL = <http://plato.stanford.edu/archives/sum2003/entries/scientific-explanation/>. **Summer 2003**

Yablo, S. (1992). "Mental Causation." *The Philosophical Review* **101**(2): pp.245-80.

Yang, Z. and D. Purves (2004). "The Statistical Structure of Natural Light Patterns Determines Perceived Light Intensity."

Proceedings of the National Academy of Science USA **101**(23): pp.8745-50.

Zeki, S. (1993). *A Vision of the Brain*. Oxford, Blackwell Scientific Publications.