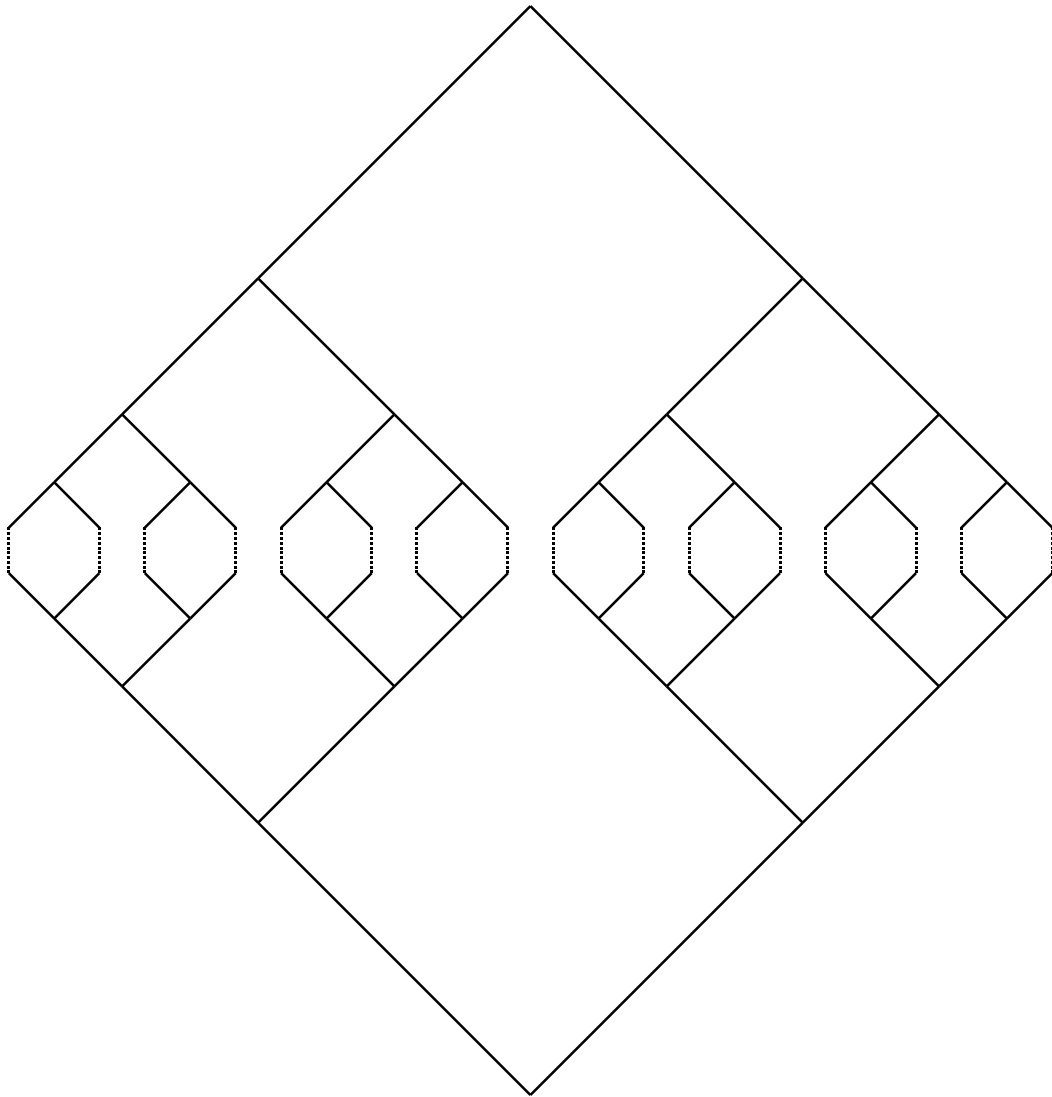


Elements of a Science of

Consciousness



Charles Fox

Elements of a Science of
Consciousness

Charles Fox

Thesis submitted for the degree of Master of Science

School of Cognitive Science
University of Edinburgh

2001

Acknowledgements

The author would like to thank the following for discussions:
Igor Alexander, Marek Binder, Gerard Blommesijn, Richard Brown,
Joseph Chen, Ron Christley, Simon Davies, John Goh,
Scott Hagen, Harutomo Hasegawa, Richard Holton, Anthony Jack,
Jenny Jack, Kieran Phipps, Barbara Piechocinska, Paavo Pylkkanen,
John Quinn, Chris Ramshaw, Antii Revonsuo, Mark Taylor,
Joseph Tolliver, Chris Town, Kevin Warwick, David Welchew,
Sarah Williams and Laura Wisewell.

The joy and the pain of writing a thesis is like giving birth;
and I thank my supervisor, Paul Schweitzer,
for his midwifery services.

This thesis was funded by an EPSRC studentship.

Work was aided by a visit to the
Toward a Science of Consciousness 2001
(*TSC2001*) conference in Skövde, Sweden.
This visit was funded by the author.

Special thanks to my family for its support over the years;
this work would have been impossible without it.

Dedication

This thesis is dedicated to the work of Douglas Hofstadter,
to which it ultimately owes its existence and inspiration.

Declaration

I declare that this thesis was composed by myself, that the work
contained herein is my own except where explicitly stated otherwise in
the text, and that this work has not been submitted for any other degree
or professional qualification except as specified.

(*Charles Fox*)

Contents

<i>Chapter 1. Introduction: Elements of a new science</i>	<i>7</i>
<i>Chapter 2. Correlates: Components of Explanation</i>	<i>14</i>
The Turing Test: a behavioral correlate.....	14
Awareness: the functional correlate.....	14
Neural Correlates.....	16
Crick & Koch: The ‘40Hz’ hypothesis	16
Edelman and Tononi: Re-entrant loops	19
Subneural Correlates.....	19
Anaesthesia: The other side of consciousness	19
Stapp: Quantum reductions as the correlate.....	22
Penrose & Hameroff: Quantum bursts.....	25
<i>Chapter 3. Phenomenology: what is to be explained.....</i>	<i>30</i>
A history of phenomenologies in Psychology	30
Introspection	31
Jamesian Introspection.....	34
Husserlian Phenomenology	36
Buddhist Phenomenology	37
Whatever happened to Phenomenology?.....	39
The Dark Ages.....	40
A new renaissance	40
<i>Chapter 4. Concepts: A new phenomenological model.....</i>	<i>46</i>
A model of concept structures	48
A model of phenomenal objects	52
A new method for phenomenology.....	57
<i>Chapter 5. Explanation: Bridging the gap</i>	<i>58</i>
From Correlation to Prediction.....	58
Nomenal and Phenomenal Realities	61
Views on the problem of consciousness	62
Compound objects exist only phenomenally	63
Rival pseudonomenologies and cultural relativism	64
Physics as pseudonomenology.....	65
Consciousness is nomenal	66
Reducing the ‘isms’	66
What is in the nomenal world?	71
Consciousness is quantum state reduction	72
Making the identity: from prediction to explanation	75
The concept of consciousness.....	78
A verificatory gap.....	79
Certainty <i>versus</i> confidence.....	80
<i>Chapter 6. Conclusion: A sketch of a theory.....</i>	<i>82</i>
Logical findings.....	82
Empirical findings	83
A research strategy proposal.....	83
Speculative findings: the best theory for <i>now</i>	85
Phenomenological simulation <i>versus</i> Artificial Consciousness.....	85
Difficult answers.....	86
<i>References.....</i>	<i>89</i>

Overview

Chapter 1. Introduction: *Elements of a new science* page 7

Consciousness is reemerging from residual dogmas of behaviorism, but with no generally-accepted framework for its study. We should work towards a practical, useful *science* of consciousness. Some ‘difficult questions’ are raised to motivate this thesis. Philosophy and experimentation should play complementary roles, providing questions and answers respectively. A good strategy is to progressively identify *transitive* correlations as a path to the *creature* correlate. The latter may be upgraded to an *identity* if particular philosophical conditions are met.

Chapter 2. Correlates: *Components of explanation* page 14

Following the above strategy, we begin by examining current candidates for objective correlates of consciousness; beginning at the highest levels of external behavior and functionality, then descending down through neural, subneural and ultimately quantum systems in the brain. We review theories concerning awareness, macroscopic γ -oscillations, pharmacological effects of anaesthetics, sub-cellular computation, and quantum effects of neural microtubules.

Chapter 3. Phenomenology: *What is to be explained* page 30

To demonstrate transitive correlation we must obtain both objective and phenomenal data. The latter is under-developed in contemporary science. However, we need not start from scratch. We draw on methods and results from traditions including Introspectionism, Husserl, Buddhism and recent Cognitivism. Key findings include: Consciousness comes in discrete *moments*; it consists of *objects*, not low-level sense-data; each object is surrounded by a semi-conscious *fringe* of associations; the *self* is simply one of the objects. A key problem: we lack a protocol for *reporting* experiences.

Chapter 4. Concepts: *A new phenomenological model* page 46

The findings of the previous chapter are integrated into a model. Its key claim is that experience consists *only* of *objects*, which are instantiations of *concepts*. Redness, triangles, faces, Pi and elements of Shakespeare plots are all concepts, and are all perceived in essentially the same way. The structure of concepts (and hence objects) is discussed, and claimed to be hierarchical, composed from lower concepts and the operations of *composition*, *generalisation* and *analogy*. The model is proposed as a basis of a new empirical method for reporting experiences, and hence for finding transitive correlates.

Chapter 5. Explanation: *Bridging the gap* page 58

The previous chapters showed how to *find* correlates. We now move to the philosophical problem of what to *do* with them. We require some kind of *identity* theory. A formalism is introduced: the *phenomenal* is the content of consciousness; the *nomenal* is the inaccessible, observer-independent ‘real’ world; the *pseudonomenal* is our phenomenal model of the nomenal. Consciousness *must* be identified with a pseudonomenon: either with a fundamental entity of physics, or a specially-positing ‘dualist’ entity. The former is more parsimonious, so likely to be true *if* we find a *pseudonomenal* correlate. *All* major current ‘isms’ reduce to the latter. Quantum identity is the only feasible way to preserve our pseudonomenology. Two objections to making such an identity – the *explanatory* and *verificatory* gaps – are discussed and disarmed. We can never be *certain* that we have the right identity, but we can be just as *confident* of it as of any other scientific identity.

Chapter 6. Conclusion: *A sketch of a theory* page 82

Consciousness must be identified with a pseudonomenon. A quantum identity would thus pose no philosophical problems *if* appropriate transitive and creature correlates could be found. If not, we are *then* forced back to substance dualism and/or a radical reformulation of our pseudonomenology. Immediate empirical research strategies are proposed for finding the correlates and for developing artificial consciousness. We return to, and answer, the ‘difficult questions’ posed in the introduction.

Elements of a Science of Consciousness

Introduction

Elements of a new science

“a nightmare to end all nightmares”
 “more frightening than being in a war”
 “being resigned to death”

- *Reports from patients who have regained consciousness
 but remained paralysed during surgery (Evans, 1987).*

This thesis may offend Cognitive scientists by asking some difficult questions; and may offend philosophers by actually answering some of them. This is not a work *within* either discipline; it is rather an attempt to go *beyond* them, to draw them together - along with other disciplines - into the rapidly re-emerging field of Consciousness Science.

During a brief period last century - roughly from 1910 to 1990 - the scientific study of consciousness became unfashionable. Prior to this, consciousness had always been the starting point for studies of the mind, at first through philosophy (e.g. Plato, 380BC; Descartes, 1641; Locke, 1690; Berkeley, 1713) then scientific introspective psychology (e.g. Brentano, 1874; James, 1890; Titchener, 1898). The unfashionable period corresponded to the dominance of behaviorism and its successor, cognitivism. The former dogma dictated that the mind was a black-box, to be studied only from a third-person perspective, in terms only of inputs and outputs. The latter dogma, influenced by the discovery of the Turing machine, allowed internal state to be added to the box, but maintained the doctrine of ‘mind as third-person machine’. During the hegemony of these establishments the subject of consciousness was held taboo by the scientific community. Searle (1992) remarked that graduate students were well-trained to ‘roll their eyes up at the ceiling and assume expressions of mild disgust’ when it was mentioned.

Cognitive Science was a success story in explaining many functions of the mind, but as its models became more detailed, so the question of where consciousness was supposed to fit in became more and more blatant, to the point where it could no longer be ignored. (cf. Kuhn, 1962.) Hence the last ten years have seen a massive resurgence of interest in the scientific study of consciousness, resulting in the launches of the *Journal of Consciousness Studies*, the *Toward a Science of Consciousness* conferences, the American Association for the Scientific Study of Consciousness, scores of consciousness books from the MIT press, Arizona University’s Centre for Consciousness Studies, and Chalmers’ seminal book, *The Conscious Mind* (1996) – seen by many as a manifesto for the movement. The University of Skövde in Sweden is due to launch the world’s first undergraduate Consciousness Studies degree in 2002.

The current state of Consciousness Science is similar to that of Chemistry just prior to the discovery of the Periodic Table. There is an electric feeling of excitement in the air that something big is about to happen – but no-one is yet quite sure what this will be. Many different ideas and approaches are emerging, from disciplines as diverse as neuroscience, quantum theory and Buddhism. But there is yet no general agreement on how to proceed – there is still everything to play for.

This thesis is my attempt to survey the contributions of these elements, and to bind them together into a sketch of a post-cognitive science of consciousness.

Some definitions

I define ‘my consciousness’ as that which is common to the experiences that I have when I am awake or dreaming, and which I do not have when I am in dreamless sleep, under anaesthesia, or unborn. I assume that other humans, including the reader, also have similar feelings and so are able to understand this definition. Once one possesses the concept ‘my consciousness’, one may then draw an analogy between oneself and others in order to obtain the concepts of their consciousnesses. By generalising over all these concepts of peoples’ consciousnesses, I define the concept ‘consciousness’, and invite the reader to do the same. (cf. Nagel, 1986.) Later in this thesis we will introduce a formal theory of concepts and give a formalised version of this definition (p.78).

This definition is based *purely* on first-person experience. It is not a functional definition. Consciousness, under this definition, has *no function*.

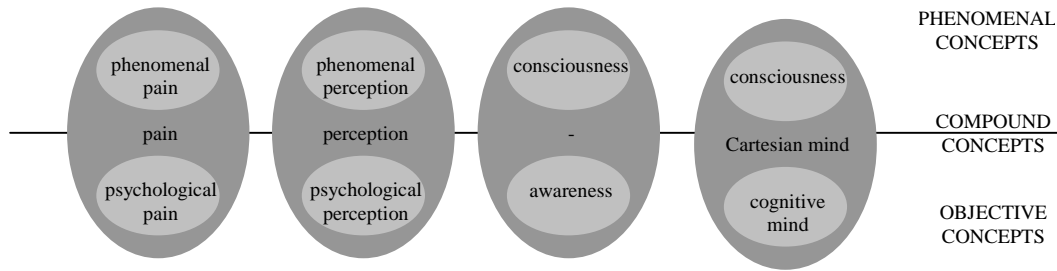
Roughly, the feeling that I have generalised in my definition of consciousness is one of the existence of a three-dimensional world, with my self and other objects in it. (cf. Revonsuo, 1995; Honderich, 2001.) I define this world as the *phenomenal* world. *Phenomenology* is the systematic study of the contents of the phenomenal world. Following contemporary convention, I also use the word ‘phenomenology’ as a synonym for ‘phenomenal world’ – we can talk about the ‘phenomenology of a person’ as we can talk of the ‘biology of a person’. (Note that ‘phenomenology’ is sometimes used to refer to a particular system of phenomenology developed by Husserl and others in continental philosophy. But here it is used in the more general sense. I will use ‘Husserlian Phenomenology’ to refer to that particular system.)

Consciousness is a *subjective* or *first-person* or *phenomenal* concept, meaning that we have used experiences which are *only* available to the subject (i.e. my personal three-dimensional world, and my self) in its definition. Concepts which are defined without using subjective components are called *objective* or *third-person* concepts, and their definitions are (in theory) equally accessible to everyone. Western science has traditionally concerned itself with providing predictive objective concepts to describe the world. Functions are objective concepts. (Objective concepts are often also called ‘physical’ concepts, but I reject this because I will later argue that Physics should be extended to include laws about consciousness.)

We observe that the presence of consciousness always seems to occur together with the presence of certain functions and other objective things. These are the functional and objective *correlates* of consciousness. Following convention (e.g. Chalmers, 1996), I call the functional correlate *awareness*. Awareness and consciousness are complementary: consciousness has phenomenology but no function; awareness has function but no phenomenology. (One might say that awareness without consciousness is blind, consciousness without awareness is empty.)¹

There is a long history of confusion about the meanings of words such as ‘consciousness’, ‘mind’, ‘perception’ and ‘awareness’. Chalmers (1996) points out that these confusions generally stem from ambiguities between words’ uses in the first and third person senses. For example, ‘pain’ can mean both a subjective phenomenal experience, or an objective psychological (i.e. functional) state. In common language there is little need to separate these two concepts, because they correlate in everyday life. There is normally no need to discuss them separately. I like to think of Chalmers’ phenomenal/functional distinction through a diagram. This shows how some pairs of phenomenal/functional concepts share common names, while we have defined consciousness and awareness to have separate names (Chalmers used phrases like ‘phenomenal pain’ and ‘psychological pain’ to distinguish such pairs):

¹ Chalmers uses ‘experience’ where I use ‘consciousness’, keeping ‘consciousness’ to mean the composite concept of awareness and experience together. However, I think that doing this is ‘to put Descartes before the horse’, since it *assumes* that awareness and experience will ultimately turn out to be the same thing. I like the word ‘consciousness’ and intend to maintain its popular meaning, as I have defined it, and I deliberately avoid a naming any compound of awareness and consciousness. (I owe the terrible Descartes joke to a *TCS2001* delegate, whom I will allow to remain anonymous.)



The word ‘mind’ has an especially strange history. Descartes conflated the phenomenal and objective concepts of consciousness and cognition into this single word. However, its meaning has now shifted to the purely functional: with Cognitive Science convention, I take ‘mind’ to mean the entire *functional* organisation of the brain. (Turing (1950) correctly predicted this shift from phenomenal to functional word meanings: “at the end of the century the use of words ... will have altered so much that one will be able to speak of machines thinking without expecting to be contradicted.”) The classic ‘mind-body’ problem should be renamed the ‘consciousness-body’ problem under contemporary definitions. This problem is more commonly known as ‘the hard problem’ in the contemporary literature.

We should note that 1970-80s ‘Philosophy of Mind’ - concerning topics such as folk psychology reduction and the ‘language of thought’ (e.g. Fodor, 1987) – meant ‘mind’ in the modern functional sense, so had nothing to do with consciousness *per se*. It was essentially just concerned with the mechanics of reverse engineering the functions of the brain. We should also be aware that some authors have used ‘consciousness’ to mean ‘awareness’: for example, under our terminology Jance’s *The Origin of Consciousness* and Dennett’s *Consciousness Explained* should be renamed *The Origin of Awareness* and *Awareness Explained*².

The aim of Consciousness Science

Now it *may* be that consciousness and awareness turn out to be the same thing, allowing us to merge the two concepts together into a single one. But there is no *a priori* reason for this. Simply to assume it would be to beg the question of the hard problem. ‘Consciousness Studies’ is often conceived as the programme of trying to find the link between the phenomenal and objective domains.

However, I would prefer to work toward a Consciousness *Science* as a *practical* and *useful* discipline. The real goal of science is not just to make conceptual links and explanations of the world, but to provide a system for making *useful predictions* about it. The concepts and explanations are just means to this end. Modern physics tells us many interesting stories about the structures and processes of atoms, for example, but its real purpose is to improve our quality of life by allowing us to design aeroplanes, televisions and microchips. Similarly, Consciousness Science may need to formulate concepts and explanations, but only as a means to the end of providing useful predictions. The predictions from Consciousness Science would perhaps be some of *the most useful knowledge imaginable* to us, concerning important situations such as how to medically resuscitate people back from the dead, whether we could achieve immortality, and how we should behave morally towards animals (and possibly also rocks, computers and corporations). More immediately, they would allow us to develop new anaesthetics and to prevent the horrific situations of conscious paralysis during surgery referred to in the quotes at the start of this chapter. In the distant future, technologies such as teleportation would also require Consciousness Science predictions about whether we can transfer a conscious mind from one place to another, allowing us to reduce the present railway slog from Edinburgh to London³ to a hassle-free instantaneous journey. (This may not be as distant as we think: Quantum Teleportation is already being taught to physics undergraduates at Imperial College of Science⁴.) A complete theory of how the phenomenal and physical are related would also allow to alter our feelings of pleasure and pain, for medical and recreational purposes. We could also imagine technologies such as direct-to-phenomenology head-up displays which could improve our personal and

² Or more cynically, ‘Consciousness Ignored’.

³ The planes aren’t much better either.

⁴ Private conversation with Imperial student. Currently only single wavelicles can be transported so there is still a long way to go.

business communication abilities. I for one would pay good money for a device which superimposed peoples' names and 'when I last met them' above their heads at cocktail parties.

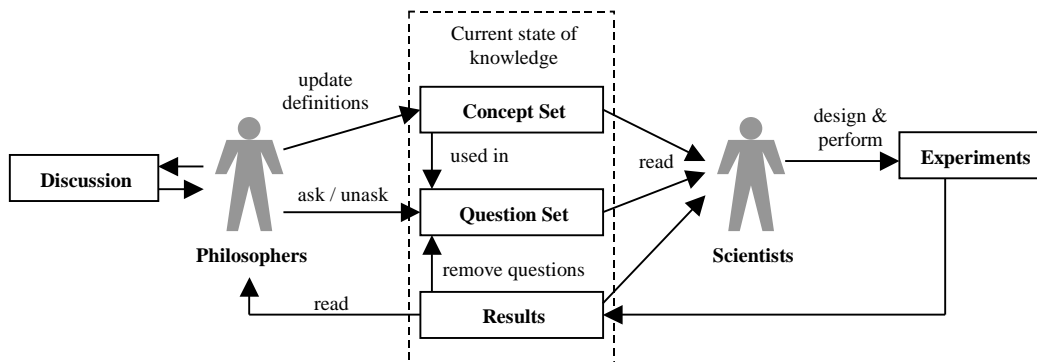
Consciousness Science is thus not just some silly academic game. It should be a practical, useful and fundable science, capable of contributing to society to improve our quality of life. (Of course it could be misused, but that is a problem for politics, not for science. Science's task is to provide the *capability* for doing good. The question of how best to *apply* capabilities is one for trained ethicists and politicians to answer.) In order to motivate this thesis, and to avoid wandering off into the interesting-but-useless discussions so beloved of philosophy, I provide below a list of 'difficult questions', (many of which are adapted from Hofstadter & Dennett, 1981) which I will try to *answer* at the end of my investigations:

- How does consciousness fit into our physical picture of the universe?
- Are other people conscious, and how can we tell?
- Which animals are conscious, and how can we tell?
- How do anaesthetics work?
- Can an appropriately programmed Turing machine be conscious, and if so how?
- Can any man-made artefact become conscious, and if so, how?
- Could we bring a recently deceased person back to consciousness, and if so, would he still be the same person?
- Could we transfer a person from one body to another, or into a machine?
- Could we build a teleporter which made an exact conscious replica of me at another location, then destroyed the original? Would it still be me at the other end? Should I agree to use the teleporter?
- What if the teleporter did not destroy the original? Which copy would be the real me?
- Can we become immortal, preserving our consciousness indefinitely?

On the complementary roles of science and philosophy

Philosophy plays an important role in the sciences, especially during their foundation and formulation. It is philosophy that clarifies our concepts and chooses questions for the science to answer. It is a trailblazer, finding a path for the heavy artillery of science to follow.

It is not the task of philosophy to *answer* questions, but to ask them. Answering them is the task of science. However, philosophy can also *un-ask* questions, by showing them to rest on incoherent or outdated concepts and assumptions. Without philosophy, science would have nothing to do – no questions to answer, no goals to work towards. The two disciples should work together in a continuing dialogue of asking and answering questions. (Philosophy without science is empty, science without philosophy is blind⁵.) As progress is made, concepts must be defined and refined so as to be useful tools for science's thinking – this again is a stock task for philosophy. The whole process forms an ongoing cycle, which progressively improves our state of knowledge⁶:



⁵ Apologies to Kant for abusing him twice in a single chapter.

⁶ Players of the popular Artificial Intelligence game *Kalah* will recognise this process of depositing in a bank during cycles.

Philosophy is often accused of making no progress, and of recycling classical ideas. But this is to miss the point: in keeping the concept and question set up to date, it may be that old definitions and questions become useful again. During their discussions, philosophers draw upon the classics, constantly asking if their insights can be reapplied in the light of contemporary scientific results. Similarly, scientists should keep aware of past experiments whose results may be useful in answering contemporary questions.

Both philosophers and scientists engage in internal, technical activities which do not *directly* contribute to the knowledge base, namely their discussions and experiments. These technical details should be kept hidden – philosophers have no need to know about experimental procedures, only their results; and scientists need not get involved in philosophical debate – only in the questions and concepts that it produces. (As Crick once said, scientists should ‘listen to their questions but not be put off by their discussions.’)

The philosopher also acts as a generalist; a ‘strategy consultant’ for the whole project. Scientists are typically skilled in narrow, specialist areas, and there is a tendency for each scientist to believe that her particular subject holds *the* key to the problem. If one’s only tool is a hammer, all problems look like nails. The philosopher is ideally placed – outside these prejudices and political hegemony struggles – to take a broader overview of the contributions of different research strategies, and to recommend new useful directions. (This is not to say that philosophy is more important, or ‘above’ science. Scientists may choose to view philosophy as ‘beneath’ their work, providing ‘foundations’ or ‘routine plumbing’ for it. Note that I have carefully drawn them side by side, as equals, in my diagram! We should also note that many individuals can, and do, take on multiple roles in both science and philosophy; I am simply drawing attention to how the different *roles* interact, not advocating a segregation of individuals into the two camps.)

Thought experiments as logical possibility generators

One of the chief philosophical methods for posing questions is through ‘logical possibility’ thought experiments. Using the current set of concepts, the philosopher constructs an potential experimental scenario, and tries to use the contents of the concepts to deduce the outcome. Many such scenarios will produce a single outcome – for example, simple physics problems like those posed as undergraduate homeworks. Given the concepts of mass, gravity and physical laws, the outcome of a scenario involving, say, a falling mass can be deduced from them. But sometimes the concepts will be incomplete, and will not be powerful enough to produce an answer. Often, the philosopher can invent several possible answers which are equally coherent with the concept set. This indicates a gap in the knowledge base; science needs to find out which of these possibilities is the correct one, and new concepts must be formulated to account for it. Before Newton’s laws, it was equally conceivable, or ‘logically possible’, that an apple might fall at any of several different speeds. Seeing this gap, Newton could then do some experiments and formulate new concepts to account for the observed result, thus removing the other logical possibilities.

Searle (1992) provides a key thought experiment of this kind, regarding the possibilities of ‘silicon brains’. Suppose that you are suffering from a disease that is gradually destroying your neurons, and that medicine has developed silicon neurons with exactly the same *functional* behavior as biological ones. Doctors gradually replace your neurons with the silicon ones. Now, there is nothing in our current science which answers the question of what happens to your phenomenology during the replacement, and there are at least two different possible outcomes which are equally conceivable under our current concepts:

- 1) There is no change – your phenomenology remains the same
- 2) You find your phenomenology shrinking. Your behavior is the same, but you find yourself losing control of it. When your left hemisphere is replaced, your right visual field disappears from your phenomenology, and you seem to lose control of your right hand – despite it continuing to pick up objects in front of it as normal: ‘You hear the doctors say, “we are holding an object in front of you; please tell us what you see.” You want to cry out, “I can’t see anything, I’m going totally blind.” But you hear your voice saying in a way that is completely out of your control, “I see a red object in front of me.”’ If the whole brain is

replaced, your phenomenology shrinks to nothing and you become completely unconscious, despite there being no change in your behavior.⁷

Searle's demonstration of the existence of these two possibilities is a wonderfully concise formulation of the central question in Consciousness Science. Our job is to find concepts (such as laws) which will enable us to predict the correct outcome of Searle's scenario and thus remove the other logical possibilities.

A cautionary note: we will later encounter a second kind of philosophical thought experiment (e.g. Block, 1978; Searle, 1980)⁸, which I think is of less utility than those of the 'possibility generation' kind. Instead of trying to *pose* questions, these experiments misguidedly try to *answer* them by a method which I call 'argument from the author's sense of absurdity.' As in possibility generation experiments, a scenario is set up, and logically possible outcomes are generated. But the author then goes on to declare that he finds one or more possibilities to be 'absurd', and that those possibilities are therefore false. However, the history of science is full of discoveries of unintuitive phenomena, which philosophers may well have thought to be 'absurd', but nonetheless were found to be empirically correct. Just because you think something is strange doesn't mean it's false. This is not to say that such arguments have no use - they do serve to draw our attention to strange consequences of our theories - but they do not *prove* anything. *Answering* possibility questions is for scientists, not philosophers.

Strategies for Consciousness Science and for this thesis

Predictive science makes progress by observing correlations and making progressively accurate predictive theories based on them. Chalmers (1997) imagines the Newton of fiction observing the apple fall from the tree. At first he might form a particular theory, 'that apple falls from trees'. Through further observations, the theory could become successively refined: 'all apples fall from trees', 'all apples fall from anywhere', 'all masses fall from anywhere', 'all masses fall where there is a gravitational field'.

The strategy of Consciousness Science should follow this pattern - and the structure of this thesis is based upon it. We should begin by noting large-scale correlations between the objective and the phenomenal, and progressively home in on finer structures. At some point, we must go beyond the mere observance of correlation, and form a predictive theory.

Chalmers (2000) distinguishes several possible meanings of 'correlate' of consciousness: the simplest being a set of minimally sufficient conditions to produce phenomenology *of some kind*; and a more complex meaning being an objective structure sufficient to produce phenomenology *containing the same structure*.

The first correlate is essentially concerned with so-called *creature consciousness* - the existence of a phenomenal world. It asks: what do we need to bring such a consciousness into being? Which parts of the brain can be removed without completely removing consciousness? Alternatively, it could be seen as specifying the minimal set of components we would need to build a conscious machine.

The second correlate is concerned with *transitive consciousness*: also known as 'consciousness of'. It takes the presence of a phenomenal world as given, and asks which objective structures give rise to the *structure* of that world. There are good reasons to believe that such corresponding structures must exist. Phenomenal worlds contain *information*, and information cannot come from nowhere - the same information must be present in the *cause* of the world. So at some level of abstraction, the objective must embody the same structure as the phenomenal. To answer the transitive question is to give a list of pairs of objective-phenomenal structural correlates. For example, activity in a particular brain area might correlate with the sensation of redness.

⁷ Searle also suggests a third possibility, in which your phenomenology remains unchanged, but your ability to put your plans and intentions into action becomes progressively reduced, culminating in fully-conscious total paralysis, like that of the anaesthetised patients mentioned earlier. However, this is a mistake - since Searle supposed that the silicon is functionally identical to the neurons, and that *only* the neurons have been replaced by silicon. So your function, and therefore behavior, must logically stay the same.

⁸ Chalmers' (1996, chapter 7) well-known 'Absent qualia, fading qualia, dancing qualia' argument also falls into this category.

We hope that tracking down the transitive correlates will lead us to the creature correlate. Once we have located the information in the objective world which matches that in the phenomenal world, we know that *something* involving that objective information embodiment is responsible for the transfer. So we will know exactly where to look for that something.

So how are we to start looking for the transitive correlates? A practical strategy is to look for large-scale approximations to the creature correlate, in the hope that they will guide us towards the information structure. (Which in turn will guide us to the creature correlate itself.)

Science has not yet found the creature or transitive correlates, but it has produced several approximations to the creature correlate. This thesis will begin by reviewing these approximate correlates, starting at the large psychological scale and digging down through physiology and eventually to physics. Hopefully this will produce some suggestions about where to look for the transitive correlates.

When we find a potential transitive correlate, we need something to correlate it *with*. We need a formal system of Phenomenology to describe the contents of consciousness. Work should begin on developing this system *now* in order that it be ready to be put into action when potential transitive correlates begin to emerge. The second part of this thesis will examine some previously suggested Phenomenologies and will draw them together with some Cognitive Science, in the third section, to produce a suggestion for such a system.

The picture on the cover of this thesis represents this overall strategy for Consciousness Science. Objective science and phenomenology are represented by the lower and upper structures. We see correlates at different layers of structure, and hope that by progressively uncovering more details, the sciences will meet in finer and finer correlates. At some point this process has to stop, and a *theory* is made (the dotted lines) which connects the correlates. The gap between the structures, to be bridged by the theory, is sometimes called the 'explanatory gap' in the literature (e.g. Levine, 1983). This tends to be used as a derogatory term, by philosophers arguing that it can never be bridged. The final part of this thesis will discuss whether bridging is possible, or whether, as Revonsuo once remarked⁹, 'we are forever doomed to march toward a science of consciousness without ever actually reaching it.' I think that it can be bridged.

⁹ During a lecture at TSC2001.

Correlates

Components of Explanation

“When I go forwards
you go backwards,
Somewhere we will meet”

- Radiohead

The Turing Test: a behavioral correlate

Before Cognitivism there was Behaviorism, which allowed discussion only of externally-observable ‘inputs and outputs’. Under this climate, Turing (1950) proposed his famous intelligence test which, under a careful reading, he also *assumes* to be a behavioral test for consciousness. We now know that Turing Test success is *not* the correlate of consciousness; but since it is the ‘highest level’ correlate ever proposed, it is still both a logical and historically interesting place from which to start our descent.

Turing is aware of the ambiguity of mental terms discussed earlier – specifically, that the word ‘think’ can mean both ‘consciousness’ and ‘awareness’. His concern is whether machines can think in the *awareness* sense, and his test is proposed as a test of this. However, in dismissing the ‘argument from consciousness’ raised in his article (i.e. a machine could be aware without being conscious, so could not be said to ‘think’ in the full sense), Turing claims that holding this objection leads to solipsism: he claims that to deny consciousness to a machine with the *same behavior* as humans would be to deny consciousness to all humans too. Turing thus assumes that behavior is the correlate of consciousness.

But claiming that ‘those who support the argument from consciousness could be persuaded to abandon it rather than be forced into the solipsist position’, is to commit the fallacy of excluded middle. There is a third option which could be taken – namely that of a *non-behavioral* correlate of consciousness. Rather than be forced into solipsism, the protagonist could take the position that the Turing Test is faulty, but that there exists a non-behavioral test which conscious things would pass and non-conscious things would fail. So she may deny consciousness to the machine, but allow it to humans. Such correlates would involve looking *inside* the test candidate, which of course was not allowed in Turing’s day.

Various arguments have been proposed to demonstrate that the Turing Test success is not an accurate correlate of consciousness. Most of these work by arguing that a machine could pass and still not be conscious (e.g. Searle, 1980; which we will discuss and diffuse later (p.68)). However, there is a much simpler reason from the ‘opposite direction’ why the correlate is not accurate. Recall the comments of the conscious-but-paralysed patients at the start of this thesis. These people would fail Turing’s test – they have no external behavior as such – but are still conscious (*ditto* for dreaming people). Hence the correlate is something other than behavior.

In the post-behaviorist world, we are now allowed to discuss internal states – both functional and neural – and it is to potential correlates of these types that we will next turn.

Awareness: the functional correlate

We have carved off, by definition, consciousness from functions. We noted, however, that consciousness appears to *correlate* with a particular kind of functioning, namely awareness. In this section we will examine the function of awareness in more detail.

Awareness appears to be the highest level of human cognition. Cognitive Science (e.g. Eysenck & Keane, 2000) frequently pictures the mind as a collection of interacting modules each concerned with different tasks. (The modules and interactions are often drawn as ‘box and arrow’ diagrams.) Typically, these modules are arranged in a hierarchy from low to high-level processing, with each level’s output feeding into the next’s inputs. At the top of the hierarchy is awareness, positioned to *integrate* highly processed and filtered information from multiple sources, such as sense data, memory and current goals. Simple, common decisions can be made and acted upon at low levels, without needed to be sent to the awareness module; complex novel decisions involving multiple modalities must be passed to awareness.

Awareness is typically held to comprise two components: a short-term memory and a problem solver. (This corresponds to the data/algorithm distinction in Computer Science.) Psychology research suggests that the capacity of the short-term memory is limited to ‘seven plus or minus two’ items (Miller, 1956) but that these items are typically compound ‘chunks’ of knowledge which can be expanded when necessary (Anderson, 1993). There is a mass of Cognitive Psychology and Artificial Intelligence research on models of problem solvers, whose frameworks include tree-searching, pattern-matching and logical systems. (See Eysenck & Keane, 2000, chapters 14-17 for review.)

Now we should be clear about exactly what it is here that is supposed to correlate with consciousness, and what kind of correlation we mean. The *presence* of both short-term memory and of the problem solver both correlate with the presence of human consciousness; but it is a well-known fact that we are not transitively conscious *of* the problem-solving mechanism – only of the contents of the short-term memory store. (If we were conscious of the problem-solver, then the part of psychology which researches its mechanisms would be trivial.) We will say that ‘awareness’ is the *presence* of short-term memory and problem solver; but that systems are aware *of* only the contents of the short-term memory. (Compare: a whole system could have ‘drunkenness’, but only be ‘drunk *of*’ the ale in its stomach. The ale could also be called ‘the contents of drunkenness’.) In functional terms, the ‘contents of awareness’ are usually defined by their reportability; and we say a module is ‘aware’ if its contents are reportable.

Most classic models of short-term memory and problem solving make no claim to being models of awareness *qua* correlate-of-consciousness – the topic still being somewhat taboo in Cognitive Science – so are not of immediate concern to us in our review of potential correlates. However, the last decade has seen a handful of functional models which *do* make the claim, and we will now examine them here. At risk of irritating repetition, it is important to stress that under our terminology, these are not yet models of *consciousness* but of *awareness*, despite some of their authors’ claims to the contrary. (To claim that consciousness is identical to awareness (or to anything) would require a major philosophical argument about why the two concepts should be equated - a topic to which we will return at the end of this thesis.)

Baars’ Global Workspace model of awareness

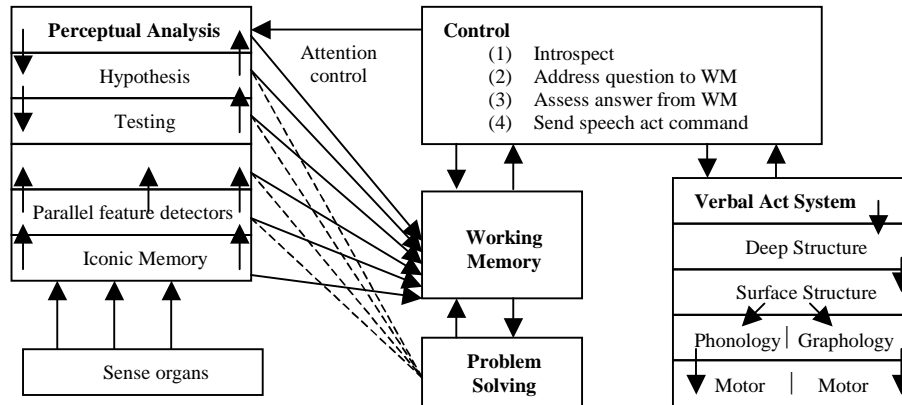
Baars (1988) draws an analogy between awareness and a large committee of experts in an auditorium. Each expert represents an independent, non-aware mind module; the whole committee represents the mind. The experts are trying to complete a task. Many of the routine subtasks have dedicated experts who can complete them individually (and new experts learn to specialise in new frequent tasks). But occasionally a novel complex subtask arises which requires many experts to collaborate on a solution. In order to achieve this, a Global Workspace, like a blackboard, is set up in the middle of the auditorium, which allows *coalitions* of experts to share data during such tasks. This blackboard is analogous to short-term memory, which is aware.

From this scenario, Baars derives several other properties of awareness. If several coalitions try to use the workspace at the same time, this may result in contradictory messages arising, and the coalitions struggling against each other to overwrite each others’ messages. Hence the tendency is toward consistent messages. This in turn leads to a tendency toward *serialisation* – coalitions take turns to solely access the workspace for lengths of time. This explains how, even though the mind is a parallel machine, the temporal structure of awareness is serial. Baars then postulates that coalitions must *compete* for access to the workspace time - by recruiting *supporters* of their access proposal, and *objectors* to rival proposals. This competitive organisation of temporality is reminiscent of how Kohonen (1982) networks self-organise *spatial* areas. The consistency requirement is also supposed to be responsible for the limited capacity of short-term memory.

The Baars model of awareness is currently being implemented by Franklin (2001), who hopes to create aware software¹ to perform useful, practical tasks such as scheduling US Navy officers.

Dennett's models of awareness

Dennett (1978, 1991) has proposed two different models of the structure of awareness. The first (1978) takes the form of a typical cognitivist box and arrow diagram:



The system is supposed to be aware of (i.e. able to report) the contents of working memory. Working memory receives information from the perceptual module, and is interrogated by a control system. The results of this interrogation are used to direct the attention of the perceptual module, and to initiate reporting acts.

Dennett's later 'Multiple Drafts' model (MDM) of awareness² (1991) controversially disposes of the notion of a single, centralised, aware module, and replaces it with a *pandemonium* system. Pandemonium simply means the presence of many individual agents competing for attention - as in the Baars model - but without a centralised workspace. Dennett borrows Baars' analogy of the collaborating experts: instead of a workspace, the experts simply send 'drafts' of their work *directly* to one another. So at any moment, there are likely to be many simultaneous and different drafts of potential solutions to the problem rather than a single one in a workspace. When a report is given, it could potentially correspond to any of these drafts, depending on which one happened to be sent to the reporting system. The contents of awareness is thus not a clearly defined set at any moment - since whether a draft is reportable is undefined until it happens to be reported. However, this appears to expose an incoherency: awareness is supposed to be the transitive correlate of consciousness; and consciousness *has* content - namely phenomenology - but awareness as given by the model does not have content. Hence they cannot be transitive correlates, contrary to the supposition that they are. So either Dennett's model is incorrect, or if it is correct it then it shows that awareness is not the correlate of consciousness. (Dennett (1991) completely ignores this issue - his book makes no mention of consciousness, only of awareness.)

Neural Correlates

Crick & Koch: The '40Hz' hypothesis

Crick and Koch (1990, 1998; Crick, 1994) propose a particular brain process as the correlate of consciousness: synchronised distributed neural oscillations in the range 30-80Hz. Formally known as *γ-oscillations*, they are conventionally referred to as '40Hz oscillations' in the literature, despite spanning the whole range from 30Hz to 80Hz. The proposed function of *γ-oscillations* is to bind

¹ Franklin confusingly calls his project '“Conscious” Software', taking great care with his scare quotes. 'Aware Software' would be a much less confusing name, since he makes absolutely no claim that the software is conscious (Private conversation).

² Dennett also confusingly calls his model the 'Multiple Drafts Model of *Consciousness*', when again 'Awareness' is meant.

together different aspects of percepts, such as color, shape and orientation, into a unitary perception. This is hoped to provide a close correlate of phenomenology.

Crick and Koch are pragmatists: their strategy is to ignore the philosophical problems and get on with finding useful neuroscientific data. Their basic assumption is that there is a single mechanism underlying consciousness, so that special (and difficult) cases such as self-consciousness and altered states are merely instances of this mechanism. These special cases can be studied later. For now, it is most profitable to choose a '*Drosophila*' case to study, and generalise from there. They make a personal, pragmatic choice to study vision.

The Binding Problem

Neuroscience and computer vision research suggest that the unconscious visual system is modular, distributed and hierarchical. Information is pre-processed in the eye, then sent to the visual cortex. Here, specialised areas extract particular features, such as color, and edges. This information is recoded to higher modules, which extract higher-level features like orientation and shape. AI theories typically suggest that this process continues upwards, until we reach levels for detecting faces, objects and people. However, while there is neuroscientific evidence, and evolutionary justification, for specialised face areas, the brain is capable of conceiving *novel* objects. Such objects could be made of *any combination of basic features*, so it would be grossly inefficient to code recognition of each possible object by a physical recognition network.

A more efficient system would be to represent objects simply by their sets of features. So rather than having a specific 'red round vertically-moving object' detector, we just have the set of detectors for red, for round, and for vertically-moving.

Our phenomenology appears as a coherent, unified view of the world. But the feature-recognition maps we have found in the brain are distinct. We perceive a single 'red round' object, not redness and roundness in the same location. Somehow, then, the features must be unified. The combinatorial problem rules out use of higher-level networks to do the unifying: somehow the parts of the feature maps *themselves* must be 'bound' to unity.

A related problem is that of ambiguity. Constructing a three-dimensional world from two-dimensional inputs is an 'ill-posed' problem: there is insufficient information in the input to produce the output. And phenomenology is not just a 3D model: it is further enriched by object-recognition's 'perceiving-as'. An identical image may be perceived as two faces or a vase. Within the perception systems there are many levels for such ambiguities to occur. AI vision systems include backwards connections in the modules hierarchy to mediate ambiguities: each layer entertains multiple possibilities, and a higher layer feeds back to choose the best one. (This is a similar process to some NLP parsers.)

It appears that the visual system contains distributed feature representations, and entertains multiple possible interpretations. But phenomenology is unambiguous, coherent and unitary. Crick and Koch suggest that the correlate of consciousness is the mechanism which takes the distributed ambiguous representations and produces the best unitary interpretation. This representation would be structurally analogous to the phenomenology.

Synchrony: a solution to the Binding Problem

A method for binding components of the feature maps was suggested by Von der Malsburg (1986), and involves a behavior of neurons which is often neglected in neural network models: *synchrony*. Most artificial neural nets are *clocked*: meaning that all the neurons effectively fire simultaneously. Typically, the state of these nets is stored, then replaced by a new state in which the new values of all the components have been calculated from the values in the previous state. But real neurons do not have this restriction – they can fire asynchronously.

Neurons from different feature maps corresponding to the same object could thus be marked by getting them to fire in synchrony. Thus a blue square and a red circle can be simultaneously represented by two different frequencies or phases of firing.

Various large-scale rhythms are found over the cortex by scalp EEG electrodes. The hypothesis is that the γ -oscillation range is used for these bindings. Note that only a small number of binding

frequencies can occur simultaneously before interfering with each other. This might explain the famous ‘seven plus or minus two’ capacity of short-term memory (Miller, 1956).

The spotlight of attention

Consciousness is pictured by Crick and Koch as a spotlight, rapidly moving around one (or possibly up to seven plus or minus two?) bound object at a time by illuminating its feature map representations with gamma oscillations. Why do we not experience phenomenology as ‘tunnel vision’ – just focussing on one object at a time? To answer this, Crick and Koch distinguish two forms of short term memory:

	<i>Duration</i>	<i>Capacity</i>	<i>Abstraction</i>
Iconic Memory	Fleeting; ~0.5s	Large	precatagorical
Working Memory	Several seconds	Limited (7±2 items)	categorical

Iconic memory is the spotlight of attention, rapidly moving from object to object, binding their low-level precatagorical features as it moves. Illuminated objects are then conceptualised and the concept representations placed in working memory – which resembles the classical cognitive model of short term memory.

Salience

How does the spotlight know what to look at? Crick and Koch postulate an additional map, corresponding to the visual field, which records the *saliency* (importance) of each location. Lesion experiments showing deficits in spatial visual attention suggest a possible location for this in the thalamus (Rafal & Posner, 1987).

Examples of experimental evidence

Gray *et al.* (1989) found 50Hz-synchronised neurons in the cat visual cortex in response to a moving bar. When the bar was replaced by two adjacent short bars, the synchronisation disappeared. This could be interpreted as the cat conceiving of the two bars separately.

Joliot *et al.* (1994) tested the generality of Crick and Koch’s theory using auditory stimuli. They played single and pairs of clicks to MEG-wired subjects, and asked them to report how many clicks they heard. (i.e. how many clicks reached their phenomenology.) For intervals less than about 14ms (corresponding to about 70Hz), only one click was perceived. The MEG recordings showed single or separated pairs of gamma activity, correlating with the phenomenology. They conclude that “binding could occur in steps or ‘quanta’ of 12-15mS” and that their results support the ‘40Hz’ hypothesis.

Normal γ -oscillation activity occurs during REM sleep but not slow wave sleep (Llinas & Pare, 1991), again suggesting a correlation with consciousness.

Can binding occur unconsciously?

Crick and Koch appear to be successful in hypothesising that gamma waves are used for binding. But the next question to ask is: Although bound concepts are required for phenomenology, is the binding process itself the correlate of consciousness? Can we have binding without consciousness? This was put to the test by Schwender *et al.* (1994), who looked for gamma activity during anaesthesia. They investigated the effects of two different kinds of anaesthetics: *receptor-binding*, which act on specific neurotransmitters; and *non-specific*, which result in a general depression of neural activity. Both types are used as surgical anaesthetics, and produce the externally observable behaviors taken by anaesthetists as signs of loss of consciousness. Whilst under the anaesthetics, patients were subjected to an auditory implicit memory task: associating the word ‘Friday’ with ‘Robinson Crusoe’.

The non-specifics blocked the γ -oscillations and these subjects showed no sign of having learned the association. This is in agreement with the Crick and Koch correlation. However, the receptor-specifics *did not block* them, and its subjects did learn the associations. The only way to save the correlation in the face of this result is to claim that the receptor-specifics did not block conscious experience, but merely blocked conscious memories from being laid down during anaesthesia. If this is true, we face the massive implication that patients operated on under these anaesthetics may be conscious of the pain during their operations; and the moral question of whether this is permissible if they don’t remember it.

Edelman and Tononi: Re-entrant loops

Edelman and Tononi (2000a, 2000b) propose a neural correlate at a similar level to Crick and Koch. They are looking for a neural process with similar properties to those of phenomenology. Their candidate process is *re-entry*. (They make the further claim that this process *is* consciousness, rather than just a correlate, but this is a philosophical question which we will examine later.)

They identify the two key properties of phenomenology as:

Integration: Worlds are unitary and coherent.

Differentiation: Billions of possible phenomenologies are possible, but they can be rapidly distinguished from one another.

Integration is preserved even in extreme pathologies, in which patients will conceive the world in spectacular contortions in order to ‘abhor holes and discontinuities’. Examples include split-brain patients seeing a full face when presented with a half-face on their visual side, and schizophrenics producing far fetched confabulations when presented with apparent contradictions in their delusions. Differentiation refers to the massive amount of information present in each moment of phenomenology, and the fact that we can rapidly make comparisons between different states; e.g. recalling memories of scenes to work out how they differ. A process implementing both of these features must be global, unifying computations from different modalities, and massively parallel, to get the speed and information bandwidth for differentiation.

Re-entry is ‘strong, rapid, parallel, recursive signalling, within functionally isolated modules’, and is Edelman and Tononi’s proposed neural correlate. It is best explained by analogy, as in (2000b): imagine four musicians, each improvising, with no conductor. Now imagine ‘a myriad fine threads’ connecting the players bodies to each other in many different places. The players go on improvising individually, but the strings tend to synchronise their movements, and produce a more integrated, coherent performance, while allowing each player to keep his own style.

A ‘Dynamic Core’ of neural areas is proposed for the brain, all intricately bound together by re-entrant connections. Neurons in the core are proposed as the mechanism necessary for consciousness. Processes occurring outside the core are unconscious. (Edelman and Tononi suggest that it is possible for the core to fragment on rare occasions, producing multiple consciousnesses in the brain – in split-brain patients, for example.)

Subneural Correlates

Anaesthesia: The other side of consciousness

While the ‘science of consciousness’ debate rages on between philosophers and cognitive scientists, practically-minded anaesthetologists have been quietly amassing a huge body of research concerning how to turn consciousness on and off for the entirely practical purpose of performing surgery. The first anaesthetic operation was carried out in 1846 by Morton, and since Langley’s discovery of receptor molecules in 1907, there has been steady progress in our understanding of their mechanisms. Surprisingly, anaesthesiology has been ignored by most consciousness researchers in cognitive science. (A notable exception is Hameroff, who we will meet in the next section.) A full understanding of the essential neural changes induced by anaesthetics would provide us with a necessary condition for human consciousness, and could point us in the direction of the creature and transitive correlates. In this section, we will review some basic pharmacological principles, and examine two current hypotheses concerning the essential mechanism of anaesthesia.

Surgeons do not worry about the philosophical problems of defining consciousness – instead they use an operational³ definition of anaesthesia, (Rang *et al.*, 1995) divided into four stages⁴:

Analgesia. Patient is aware but drowsy. (Degree depends on anaesthetic concentration)

Excitement. Loses awareness; will not respond to nonpainful stimuli. May move and talk incoherently, and show cardiovascular irregularities.

Surgical Anaesthesia. Movement ceases; respiration regularises. Reflexes lost. Muscles relax.

Medullary Depression. Respiration and vasomotor control cease, leading to death.

There are many known types of anaesthetic, each with its own combination of additional side-effects.

What kind of targets do anaesthetics affect?

Until the 1990s, the dominant view was that anaesthetic molecules were very non-selective of action targets, and acted by reducing general neural activity. This was supposed to result from the anaesthetic molecules dissolving into fatty parts of neuron cell membranes ('lipid bilayers') so as to reduce their functioning – perhaps by reducing ion channel opening. This view was built on Meyer and Overton's experiments in the early twentieth century, which showed a correlation between anaesthetic potency and fat solubility. However, Franks and Lieb (1982, 1984) found exceptions to the Meyer-Overton correlation, and proposed that anaesthetics act much more selectively - by binding directly to specific proteins, such as neurotransmitter receptors. Different anaesthetics bind to different types of target, and some bind to multiple types of target. Research has since focussed on identifying which of these specific targets is ultimately responsible for loss of consciousness, and which are merely responsible for side-effects. To follow the debates, we must first review a little basic neuropharmacology (Rang *et al.* 1995; Purves *et al.* 1997):

Some basic neuropharmacology

The protein targets we are concerned with are neurotransmitter receptors in synapses. Synapses join neurons together to allow them to communicate. When the presynaptic axon's electrical pulse reaches the synapse, it causes one or more chemical neurotransmitters to be released from the end of the axon. The dendrite of the postsynaptic neuron contains sections of protein which these neurotransmitters may bind to – these are *receptors*. Once activated in this way, the receptors cause ion channels to open in the dendrite, which produce (or inhibit) an electrical pulse. In this way, electrical communication is passed between neurons.

There are over 100 types of neurotransmitter, but here we will be concerned with just four. γ -aminobutyric acid (GABA), Glutamate, Acetylcholine (ACh), and Serotonin (5-HT). Each type of transmitter can be received by multiple types of receptor. There are two broad classes of receptor: *ionotrophics* are ligand-gated ion channels – they themselves are ion channels which open when bound to by the transmitter. *Metabotrophics* are slower in action and more indirect: binding causes the receptor to release *G-proteins*, which then bind to an *effector-protein*, which in turn causes a nearby ion channel to open. Ionotropic receptors which concern us include the glutamatergics (meaning that they receive the neurotransmitter Glutamate): NMDA and AMPA - and a GABAergic receptor called GABA_A. GABA is also received by a metabotropic receptor called GABA_B. ACh is received by a further family of ionotrophics.

Anaesthetics are a form of drug. Drugs are substances artificially introduced to the body, which bind to protein targets usually used by natural biological processes, so as to change the functioning of these processes. The protein targets for anaesthetics are receptors. There are two basic ways in which a drug can affect a receptor: *agonists* bind to the receptor, mimicking the effect of the transmitter it is designed to receive. *Antagonists* act by blocking the receptor, preventing it from detecting the transmitters and so disabling it.

³ No pun intended.

⁴ The author was somewhat disappointed when he finally got to meet a professional anaesthetist for lunch, and asked her what her profession had to say about the Nature of Consciousness. Dismissing the question as something of an irrelevance, she said tersely that anaesthetics just *work* – she didn't care how or why as long as they were useful for surgery – and went back to drinking her wine.

The GABA hypothesis

General anaesthetics can affect a large number of molecules in various ways, especially if administered in high enough concentrations. The problem is to distinguish which ones are responsible for loss of consciousness. Franks and Lieb (1998, 2000) consider three criteria in their *in vitro* search for the correlating target:

Sensitivity. Preference should be given to those putative targets whose behavior is significantly affected at the *same level of concentration* as required for surgical anaesthesia.

Stereoselectivity. Many anaesthetic molecules have *optical isomers* (different configurations of the same molecules, such as mirror-image structures) which have different anaesthetic effects. For example, the anaesthetic etomidate is 15 times more powerful in one configuration than in its mirror image. (Tomlin *et al.* 1998). Preference is given to putative targets whose behaviors show similar stereoselectivity to those of the anaesthetic effect.

Nonanaesthetics. If putative receptors are also activated by chemicals which are known to have no anaesthetic effect, then they can be ruled out.

Using these criteria for a wide range of different anaesthetics, they conclude (Franks & Lieb, 1998) that effective receptors are a set of GABA, 5-HT and ACh receptors. Interestingly, these form a closely related genetic family. The GABA_A receptor in particular is the best *in vitro* correlate of behavioural anaesthetic effects. Could the activity of this receptor family be the correlate of consciousness? Sadly, no. Despite having the best correlation of all receptors, for a wide range of anaesthetics, it is not a perfect correlate: there is a particular anaesthetic, *ketamine*, which does not affect the family. So there must be more to anaesthesia than deactivating the GABA family.

The NMDA hypothesis

Flohr (2000) has put forward an alternative receptor correlate of consciousness: the NMDA receptor. (See picture on next page.) His work begins with ketamine, the rogue anaesthetic which defies the GABA hypothesis. Ketamine is a relatively new anaesthetic, discovered in 1966, and produces notably different side-effects from 'classic' anaesthetics: a trancelike state, abnormal perceptions, hallucinations, and ego disorders. Flohr thinks that these side effects act 'closer to consciousness' than those of other anaesthetics, showing ketamine to be an especially pure anaesthetic. He thus originally hypothesised that studying its mechanisms would reveal a small, pure set of affected receptors which are most fundamental to consciousness. The NMDA receptor is the only known major target for ketamine (Franks & Lieb, 1998) – leading Flohr to conclusion that it is this, and not the GABA family, that is the correlate of consciousness.

But how do we explain Franks and Lieb's results that the GABA receptor family, and not NMDA receptors, correlate with almost all anaesthetics except ketamine? The answer is twofold: First, their experiments were done *in vitro* – they tested the effects of the anaesthetics on the receptors in isolation from the rest of their natural environment. Secondly, Flohr (2000) notes that *in vivo*, NMDA receptors always have GABA receptors nearby. His hypothesis, then, is that most anaesthetics all act *indirectly* upon the NMDA receptor, often via the GABA_A. Ketamine is unusually pure in that it affects the NMDA directly, and hence gives the purest form of anaesthesia.

Flohr (2000) reviews a large body of evidence from NMDA experiments, and concludes that:

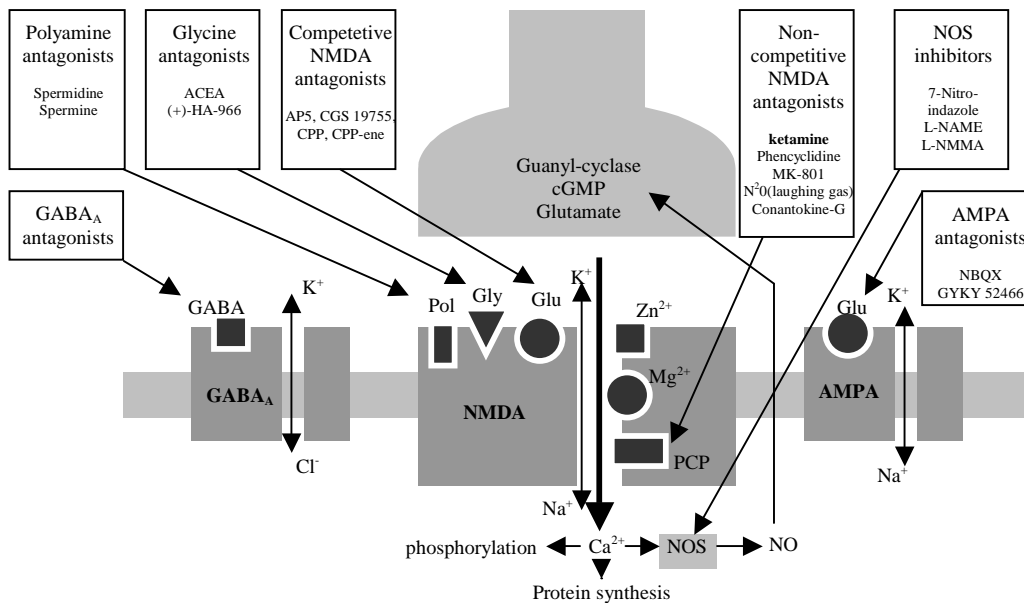
- (1) Normal functioning of NMDA receptors is *necessary* for consciousness
- (2) All other brain processes in the absence of NMDA receptors are *insufficient* for it.
- (3) Restoring NMDA functionality after (2) is *sufficient* to restore consciousness
- (4) Therefore, NMDA functionality is *the* correlate of consciousness.

(In fact, he draws the stronger conclusion that NMDA functionality *is* consciousness, but we are currently concerned only with finding a correlate. The philosophy comes later.)

Flohr identifies potential binding sites on and around the NMDA receptor for *all* major known anaesthetics (see diagram below), and impressively, predicts the anaesthetic effects of other molecules which would affect the receptor. For example, AMPA receptors, like GABA_A have similar knock-on effects on the NMDA receptor, so AMPA antagonists are predicted to be anaesthetics. This is indeed the case (McFarlane *et al.* 1992). Furthermore, it has recently been shown (Jevtovic-Todorovic *et al.*,

1998) that laughing gas (nitrous oxide) works by affecting NMDA receptors directly – and this would explain the fact that its behavioural effects are very similar to that of ketamine.

What could be so special about NMDA receptors that makes them responsible for consciousness? They are found almost exclusively in the cortex and hippocampus – areas usually associated with consciousness and memory respectively. Furthermore, they have an activation mechanism unique among receptors: most receptors are activated by the solely presence of a chemical (a neurotransmitter) – but the NMDA has the *additional* requirement of being at a particular electrical potential for activation to occur. Flohr (2000) shows that this dual activation requirement is theoretically sufficient as a detector of pre- and post-synaptic neuron firing coincidences – and thus for Hebbian learning to occur. (Hebb, 1959). In fact, the NMDA receptor is capable of both long and short term Hebbian learning – suggesting roles in both long and short term memory. And short term memory brings us close to consciousness.



*The NMDA receptor and its neighbours, showing targets of major and lesser-known anaesthetics.
(redrawn from Flohr, 2000)*

An attack on the NMDA hypothesis

GABA researchers Franks and Lieb (2000) retaliate against the NMDA hypothesis as follows: If NMDA receptors are the ultimate target of all anaesthetic action, and if they are responsible for hallucinogenic side effects, then surely all anaesthetics, in ultimately affecting the NMDA receptor, would bring about the same side effects. But they don't. Ways of explaining this fact away would involve tacking on "special pleading" postulates to the NMDA hypothesis, which would make it more complex and so less appealing. Nevertheless, they conclude that making and testing such postulates will be a fruitful area of scientific research to further test the NMDA hypothesis.

Stapp: Quantum reductions as the correlate

"Anyone who claims to understand quantum mechanics is either lying or crazy."

- Richard Feynman

The consciousness research we have reviewed so far is mostly the work of 'traditional' Cognitive Scientists: Psychologists, Neuroscientists and Philosophers. All are struggling to find a place for

consciousness in the otherwise physical world. Most of them seem to have an implicit Newtonian picture of the physical world: in which increasingly complex structures are ultimately constructed from small, mechanical, deterministic particles in 3-dimensional space. The ontology of physics is assumed to comprise a set of fundamental particles, a spacetime in which they exist, and a set of rules determining how they interact. And this ontology does not contain consciousness.

However, Stapp (1996) points out that this picture of physics is wildly out of date, and what these people have missed is the surprising fact that all contemporary interpretations of physics are essentially *dualistic*. They all contain two components, very roughly corresponding to a mathematical model of the universe, and a conscious observer.

This dramatic addition of the observer is due to Quantum theory, which models particles as existing over a spread of locations and states instead of at single points and single states. Such particles are said to be in a *superposition* of states. This gives correct predictions for many situations in which classical physics fails, so is our currently accepted model of reality. The predictions are deterministic as long as we picture the universe as always existing in superpositions: they predict exactly how a superposition state evolves over time into future superposition states. However, this is not how our phenomenological worlds appear: we don't see superpositions but single, coherent states. Somehow, the superposition *reduces* into a single state.

Current physics is unable to predict which state will be selected, and there is not yet any agreement about how, or when, reduction occurs. The best it can do is give a set of probabilities for each state occurring. (But randomness in science is merely a synonym for 'we don't know how to predict this'; multiple logical possibilities are open under our current scientific conception of such situations, waiting to be closed by new concepts.) We will now follow Stapp in examining the four major interpretations of the quantum model, each of which tells a story about how reduction occurs. Essentially, it is the process of reduction that provides an extra component in the ontology, which Stapp proposes to be associated with consciousness.

The Copenhagen interpretation: mathematical fiction

Bohr (1934) advocated a massive philosophical shift for physics: that it should give up its high goal of disclosing the real essence of the universe, and instead should merely 'track down as far as possible relations between the multifold aspects of our experience.' That is to admit that the entities postulated by physics may be completely fictitious, yet they are useful for making practical predictions about the world. The quantum superposition model is just a mathematical device for making predictions about our experiences – we can never know the true nature of reality. (This is what philosophers have been stressing for centuries). Reduction and its state probabilities are not real processes; they are just part of the model. So this story contains two components: conscious phenomenology is given pride of place, and is accompanied by a mathematical device for predicting its contents.

The Bohm interpretation: waves and particles

The Copenhagen interpretation can be seen as just giving up on finding a realist interpretation. Bohm (1952) attempts to rescue physics' metaphysical status by postulating a two-part ontology: each quantum entity contains both a probability wave and a particle which 'rides' upon it (like a surfer). The wave is sufficient for performing deterministic superposition calculations, but our experience is only of the particle, which we find at a probabilistic point in the wave. In this story, we can never experience the wave directly, we can only infer it from the particles which we experience. Note that without a conscious observer, the waves are sufficient to describe the universe, always existing in superpositions. But since we don't consciously experience superpositions, we need to postulate the particle as a means of bridging the gap between reality and our experience.

The Heisenberg interpretation: real random reduction

A problem with the Bohm story is that all the parts of the wave which we didn't experience continue to exist and evolve after our experience. This results in the universe largely comprising of invisible superpositions, which is not a very parsimonious explanation of the world that we see. Heisenberg (1958) wants to cull these superpositions by proposing real, random Reduction (**R**). In this story, our phenomenology *does* show us the complete world (rather than a randomly chosen particle projection of the superposition) – because the superpositions really do reduce into a single reality. Superpositions exist for some length of time, then somehow are reduced to one of their component states. However,

there is nothing in our current physical model that predicts when **R** occurs. We know that at least when we are looking, **R** must occur at least as rapidly as our phenomenology is updated, because we never *see* superpositions. (We will later see (p.39) that the update frequency is about 50Hz – roughly the same as γ -oscillations.) But other than that, physics is silent as to how or why **R** occurs. Yet again, we must postulate something dual to the physical world to explain it. Furthermore, it seems impossible to make non-subjective measures of when the reduction occurs: for if we had such a measuring device, that device *itself* could just go into a superposition of measurements for the whole period up until we looked at it and became sure that reduction had occurred. (This is reminiscent of Berkeley's idealism). Given this impossibility, a parsimonious way to extend our model of physics is to claim that **R** occurs precisely when a consciousness looks at a superposition: consciousness *causes* 'subjective reduction' (**SR**) (Wigner, 1961; von Neumann, 1932). So again we have added a consciousness dualist element to physics.

The Everett interpretation: many minds

Incorrectly referred to as the 'many worlds' theory by popular science, Everett's (1957) interpretation goes back to Bohm in postulating the continuing, non-reduced existence of the universe of superpositions. Recall that Bohm's theory was problematic because it postulates the existence of many alternative states which are inaccessible to the conscious observer after an observation has shown him to be in a particular state. The other states are still there for ever, but seem 'wasted' since no-one ever sees them. Everett's story is that the conscious observer *himself* goes into superposition when he makes an observation. Now from the point of a particular observer *in a particular state*, it *appears* that a random objective reduction has taken place. And the continued existence of the other (now inaccessible) states, each containing the same observer in a different state, makes it look as though there are 'parallel universes' of different **R** outcomes. But taking a God's-eye 'view from Nowhere', the universe is still singular, deterministic, and made of superpositions. It is the consciousness that has been changed (i.e. superposed) by observation – the rest of the universe is unchanged. **R** is an observer-relative illusion. To ask 'but why am *I* in *this* state' is a non-question - like asking 'why is now *now*?': *given* any observer in a particular state, he asks that question about his own state. His counterpart in another state (from the same superposition) may well ask the same question about his own state.⁵

Stapp notes that at first glance this picture seems to remove the need for the now-expected dualist element, since no **R** actually takes place. However, he then points out that the whole point of quantum theory is its to make *probabilistic* predictions about the world: so just allowing for there to be two future subjective observer states is not enough – Everett must give a story to explain why one state appears more likely than the other. Somehow, each superposed state must have a probabilistic weighting for what Stapp calls the 'subjective probability to occur'. His argument here is a little unclear (Stapp, 1996, p.204), but he suggests that this theoretical mechanism would essentially form the dualist element of the story.

Macro collapse as correlate

Each of the interpretations must thus contain a dualist element relating the superposed physical world to the reduced-state phenomenological world. The Copenhagen story is not about the real world, and is basically just giving in. Bohm's story is workable, but at the expense of being grossly unparsimonious (by postulating almost infinitely many more invisible states in the universe than those containing its

⁵ An interesting philosophical conclusion seems to result from this story, which occurred to me many years ago (inspired by Hofstadter, 1985), and which I finally get the excuse to write down: Under the Everett interpretation, one's own death is impossible. Since for any situation where my death may occur, such as being executed by firing squad, there must always be some probability, no matter how miniscule, that death does not occur, e.g. the firing squad misses. My body thus goes into a superposition of being alive and dead simultaneously (from a God's eye view). Now, for most superposition decisions, there is a subjective illusion of a random **OR**, and my consciousness appears to move in to one or other of the next possible states at random. But since my consciousness only *exists* in the 'alive' state after the potential-death event, there is only that future state for my present consciousness to appear to move into. To the outside observer (maybe a member of the firing squad, who successfully executes many people each day), the probability of my death is high, and in most future superposition states he will see me dead. But from *my* point of view, the chances of survival are 100%, since I (my consciousness) only exist in those worlds where I am alive to make the comparisons. Since there is always some small possibility of escaping death in any particular situation, the strange conclusion is that everyone is immortal from their own point of view. (My Christian friend points out that this argument rests on the assumption that bodily death kills consciousness; and it would not work if consciousness could escape to an afterlife. It is perhaps simpler for Everettians to grant the existence of this afterlife rather than accept my bizarre immortality conclusion!)

conscious observers). The Everett story is incomplete, lacking a probabilistic mechanism. So Stapp comes out in favour of the Heisenberg Reduction version, with the Wigner-von Neumann dualism of consciousness as the cause of subjective reduction. The reason **SR** is the correlate of moments of consciousness is because **SR** is postulated precisely to add predictions to our physics which concern features of our conscious experience (i.e. the singularity of phenomenology). If it wasn't for the existence of consciousness, reduction would not be needed as a fundamental entity in our physics.

Recall that the most parsimonious frequency for reduction events is such that one occurs every time we look; every time we have a moment of consciousness. (Recall that phenomenology moments occur about every 50Hz). So Stapp suggests that superpositions might grow to a macroscopic size, spreading from the world into parts of the brain, until collapsing into a moment of phenomenology. He doesn't imply that consciousness can control the *outcome* of the **SR** – just that it correlates with **SR**. The outcome is still random. He tentatively suggests that this may be a useful feature for systems to evolve, since by collapsing the superposition, they simplify the world and make it easier to predict.

Summary

Stapp has shown that contrary to the implicit assumptions of most Cognitive Scientists in search of a place for consciousness, all contemporary interpretations of physics *already* have such dualist elements. Although there is still controversy over which interpretation is correct, Stapp suggests that the most parsimonious story is that of Subjective Reduction (**SR**), in which superpositions exist for short periods of time before objectively collapsing into a random component state. Current physics has not agreed on a mechanism for reduction, and Berkelian problems appear to make gathering objective empirical evidence impossible. So Stapp suggests that the simplest solution is to equate the **SR** action with moments of consciousness. Superpositions reach macroscopic size in the brain, perhaps corresponding to simultaneous superposed preconscious phenomenologies, before being reduced at about 50Hz.

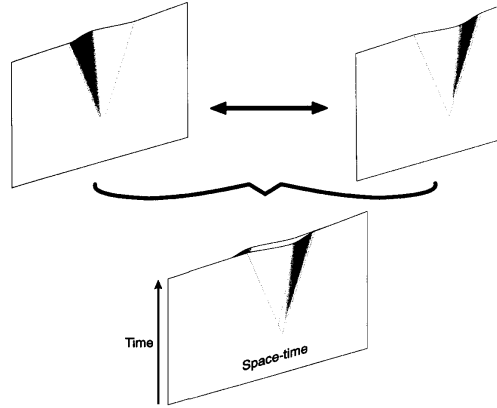
Penrose & Hameroff: Quantum bursts

Penrose and Hameroff (e.g. Penrose, 1994; Hameroff & Penrose, 1996, Hameroff, 1998a) have developed an alternative theory to account for both Reduction and consciousness. Their theory has a controversial history, and has grown substantially since its first incarnation was published (Penrose, 1989). Hence much of the criticism launched against it (most notably by Grush & Churchland, 1995) no longer applies to its current incarnation (Penrose & Hameroff, 1995). Here we will describe only the current version of the theory.

Objective reduction: bubbles in spacetime

As we noted earlier, there is something missing in our physics: namely, a mechanism for reduction. The reduction process reduces a superposition of states to single component state, seemingly selected at random. We need a theory to explain when this occurs, and to explain which state is selected. Penrose's answer is that superpositions, like growing bubbles, can only sustain themselves to a certain size before 'bursting' – or *self-reducing*. This process is entirely objective, so is named Objective Reduction, **OR**. It is held to be deterministic, but non-computable, due to the effects of 'hidden variables' in superposed reality whose values are inaccessible to us. (Note that the postulating of these variables would be seen by Stapp as the dualist element required to connect superpositions with single-state phenomenology.)

Einstein showed that spacetime is curved, and that its geometry is altered by the positions of masses. The curvature is gravity. Spacetime is like a rubber sheet: if you place a heavy object on it, it stretches. So two spacetimes with different configurations of masses will have different geometries. This presents a problem for quantum theory. When an area of spacetime goes into superposition, we are supposed to get two states with different mass configurations sharing the *same* piece of spacetime. But we now see that because two different geometries are created, spacetime *itself* must go into superposition. As spacetimes have no global coordinate system, it becomes impossible even to say which point in one spacetime corresponds to a point in the other. Over time, the size of the superposed space grows, since the contents of the superposed area gradually affect nearby areas. Penrose often uses the diagram on the next page (taken from Hameroff & Penrose, 1996) to show the two superposed spacetimes growing bigger and further apart:



Physics is still unclear about such *Quantum Gravity* effects – and in fact the basic principles of relativity come into conflict with those of quantum theory if the bubbles reach cosmological scales. What Physics would like is a theory to burst these bubbles before they get too big – an **OR** theory. Penrose suggests that **OR** might occur when a bubble reaches a certain size threshold, and is unable to sustain itself.

Now, remember that these are bubbles in *spacetime*, not just in space. So the ‘size’ of a bubble depends on both its age and its volume. In order to self-reduce (burst), a bubble can either be small for a long time or big for a short time (or a mixture of both). Penrose has calculated the amount of time, T , required for a bubble of a certain size to self-reduce (where E is a measure of the energy between the mass distributions of the two superposed states – i.e. a measure of the ‘difference’ between them; and h is Planck’s constant):

$$T = \frac{h}{2\pi} E^{-1}$$

In addition to *self-reducing* (bursting), bubbles can also be reduced by interference from the environment. The masses in each state must maintain *coherence*, which is extremely delicate and easily disturbed by interactions with other particles. Coherence is a reason why quantum computers are difficult to build: the coherent elements must be well *insulated* from their environment. Hameroff and Penrose claim that whilst *self-reducing* is deterministic (though non-computable), an externally-caused reduction is still *random*. (e.g. Hameroff, 1998a: “these states are chosen randomly”).

(There still appears to be some disagreement between Penrose, Hameroff and their co-workers about the exact statuses of the random/non-computable processes. Hameroff’s collaborator, Scott Hagen (private conversation) explains that the theory is very much ‘in-progress’ and different people have different ideas about these details.)

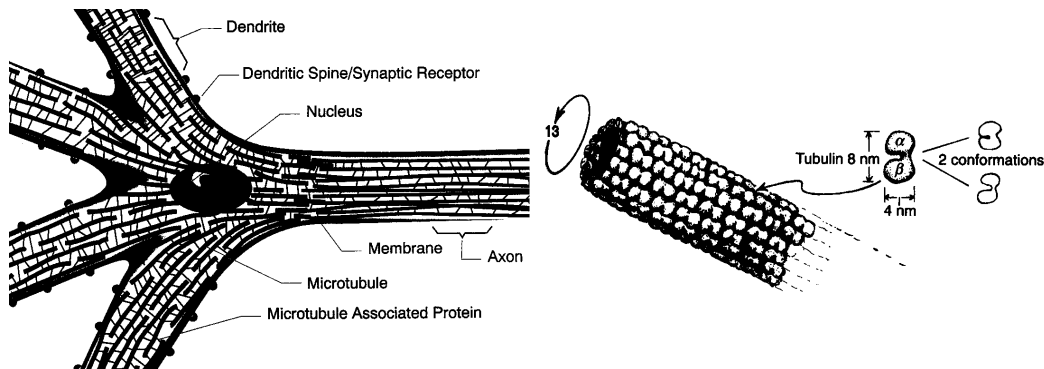
Under this relation of space and time requirements, a single electron would have to remain in superposition for longer than the age of the universe in order to self-reduce. Larger objects would take less time, but become more delicate to interference (since they consist of many particles all of which must maintain coherence with each other). In short, meeting the conditions for self-reduction is very difficult and would require specialised insulation mechanisms. Self-reduction is rare in the universe.

A brain-sized object would be required to maintain superposition for the order of tens or hundreds of milliseconds. If the brain featured an appropriate insulation mechanism, this would allow superpositions to form and self-reduce at roughly 50Hz. Penrose and Hameroff hypothesise that the brain does in fact contain such processes, and that the each burst is the correlate of a moment of phenomenology.

Microtubules

Hameroff suggests that the brain’s specialised equipment for orchestrating quantum reduction is to be found in neural *microtubules*. Microtubules are long hollow cylinders which form much of the cytoskeleton. They are the scaffolding which holds the cell in shape, along with microtubule-

associated proteins (MAPs) which connect them together. They are also used to guide the movement of neurotransmitter components along axons from the nucleus to the synapses. Neurons have an exceptionally large number of microtubules: around 10^7 inside each⁶, and unlike all other types of cell, they are arranged linearly, not radially. Each microtubule is built from *tubulins*, arranged in a hexagonal lattice. A circle of 13 tubulins form each layer of the microtubule. Each tubulin can exist in one of two states, determined by the actions of a single component electron.



Left: Schematic diagram of a neuron, showing microtubules interconnected by MAPs.
Right: Microtubule structure, showing individual tubulins. (from Hameroff & Penrose, 1996)

The closed-tube nature of microtubules makes them ideal for insulating the coherent quantum states required for self-reduction. They are also at the right scale to mediate between the quantum level of electrons and the neural computation level.

In heated opposition to the Churchlands (who claim that treating neurons as black boxes is sufficient to explain the functioning of the brain), Hameroff (e.g. 1999) says that the sum-threshold neuron model used by neurofunctionalists is “a cartoon, skin-deep portrayal that simulates a real neuron as much as an inflatable doll simulates a real person” and that “apparent randomness exists at all scales of the nervous system” which needs to be non-randomly accounted for. Specifically, he notes that only 15% of axonal potential spikes reaching the synapse actually cause the release of neurotransmitters. Something more detailed must be influencing them. This ‘something’ is the microtubule activity.

Quantum computation

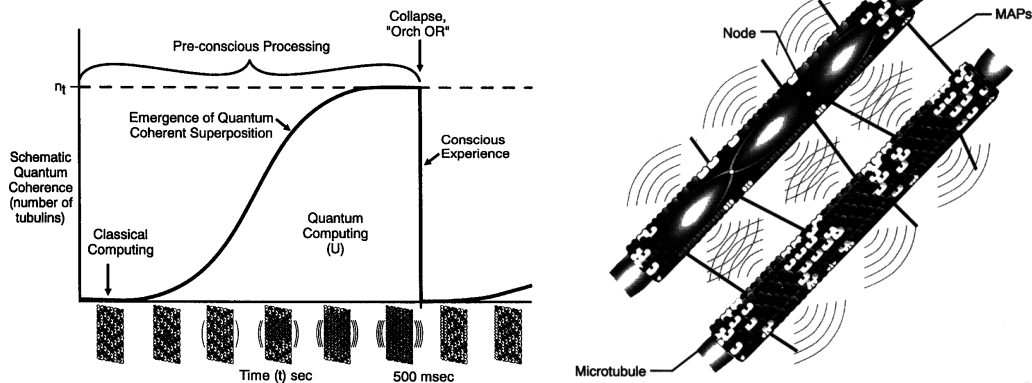
It is a historical fact that theories of the brain tend to be couched in terms of the latest man-made technology. Such thinking has previously been influenced by clockwork, serial computers, and the (parallel) internet. The cutting edge of theoretical computer science today is *quantum computing*, which in addition to 0 and 1, allows information to enter a third, superposed state of being both bits at the same time. This quantum equivalent of the bit is called a *qubit*. (For an introduction, see Nielsen & Chuang, 2000) Following tradition, Penrose and Hameroff draw a new technology analogy. Microtubules are used to perform quantum computation, and the two possible states of their tubulins represent qubits. This extra computational power is responsible for the apparent randomness of neural firing.

There is already good evidence (reviewed by Bray, 1995) that subcellular proteins can and do perform classical (non-quantum) computation. *Paramecia* is a single cell bacterium, yet is still capable of swimming, finding food and mating – all without neurons. Bray shows how proteins can perform Boolean operations, fuzzy logic, and mimic the McCulloch-Pitts neuron. (He further suggests that large networks of proteins would have similar properties to neural networks - this casts serious doubts on the Churchlands’ neurofunctionalism.)

Penrose and Hameroff propose a three-phase cycle of (super-Turing) computation in microtubules. Recall that each tubulin can be in one of two possible states, or both. The first phase is classical, with tubulins in non-superposed states performing protein computations. Superposition then begins to emerge in a small area of tubulins, and it spreads through interactions with nearby areas. During this

⁶ Compare this with the 10^{12} neurons and 10^{15} synapses in the brain

second phase, powerful quantum computation is taking place. Finally, when the superposition reaches the threshold spacetime size, it objectively reduces. The tubulins are now in new non-superposed states, and their configurations are uncomputable by Turing machine.



Left: Time diagram showing phases of classical and quantum computation, then reduction. The lower pictures show the superposition spreading across the surface of the microtubules. The reduction is the consciousness correlate. Right: MAPs create harmonic nodes to orchestrate quantum waves, 'tuning' the outcome probabilities. The upper tubule is cross-sectioned to show the waves inside; the lower one shows the external tubulins in qubit states.

How exactly does the reduction take place? Coherent photons are generated by the tubulins surface, and are insulated inside the microtubule. Because of their tube shapes, microtubules allow the photons to behave as standing waves: a little like sound waves in a musical pipe. And as a musical pipe can be tuned by setting up nodes to create harmonics of the wave, so microtubule-assisting proteins attach to the microtubule to tune the harmonics of the quantum oscillation. This tuning influences the possibilities and probabilities for the objective reduction. This 'orchestrated' objective reduction is termed **Orch-OR** by Penrose and Hameroff. Hameroff (1998) compares **Orch-OR** to a sailor steering through complex vortices of wind. The wind is analogous to the values of the 'hidden quantum variables' which determine the outcome of the reduction – it is deterministic, but non-computable⁷. However, the sailor can influence his chances of certain movements by tuning his sail. The sailor might not arrive at the exact location he wishes for, but he can get close to it.

We have so far seen how quantum coherence can arise in single microtubules. But the crux of the theory is that the whole volume of brain responsible for consciousness goes into a single coherent superposition and reduction. How could this happen? We need mechanisms for tubule coherency *within* neurons, and for coherency *between* neurons. The feasibility of retention of coherency *within* cytoplasm has already been shown by Walleczek (1995). Coherence *between* neurons is more difficult to explain, though Hameroff (1998a) suggests that quantum tunnelling through electrical gap junctions (non-chemical synapses, ignored by neurofunctionalist models) may be a possibility. Recent research (Richter *et al.*, 2000), originally intended to improve MRI contrast, has demonstrated that coherence can be sustained in the brain over distances as large as 1mm. (However, this involved artificially inducing the coherency – so only demonstrates that natural coherency is *possible*, not that it definitely exists.)

Allowing the whole brain (or rather, the part of it responsible for activities correlating with consciousness) to superpose coherently gives us a mechanism for Stapp-like processing: the whole preconscious phenomenology can superpose. Multiple future phenomenologies could be entertained and appraised, then **Orch-OR** could fix probabilities to 'sail towards' the best outcome.

Attractive features

The Penrose-Hameroff model has been accused of being highly speculative, but its risk is balanced by an array of neat explanations which are unavailable to other models. First, the determinate but non-computable nature of consciousness rescues our feeling of free will (we will discuss this in more detail

⁷ The analogy is not perfect: the wind would be computable *in theory*, though impossible in practice. Reduction, in contrast, is uncomputable even in theory.

later, see p.75). Secondly, it allows us to *quantifiably* calculate how conscious other animals are, using the brain size / coherence time formula. For example, the worm *C elegans* would have about 2 moments per second compared to the human's 50. This would be very attractive to utilitarian animal ethicists. (Is Singer (1979) correct to prefer killing a newborn baby to a cow? Just work out the numbers!). Third, the model can be used to identify when consciousness arose in evolution: what was the first creature capable of producing a single self-reduction? Hameroff (2001) suggests that urchins and worms 540 millions years ago are likely candidates. (This coincides with the 'Cambrian Explosion' – suggesting that the arising of consciousness's increased computing power contributed to accelerated evolution.) Fourth, the brain-scale macroscopic coherency provides an explanation for *binding* – and its roughly 40Hz cycle fits gives a underlying explanation for the Crick and Koch correlate. Finally, Hameroff (1998b) has applied the theory to develop a detailed explanation of anaesthetic action. We earlier showed that anaesthetics bind to proteins, and the NMDA appears to be the vital protein site. But Hameroff provides a further level of explanation about what happens next: the presence of the anaesthetic prevents tubulins from going into superposition, and so disables self-reduction. (There are also deep philosophical advantages of the Penrose-Hameroff model, which we will consider later.)

Criticisms

The Penrose-Hameroff model has been barraged by criticism from the Cognitive Science establishment (especially the Churchlands) who perhaps feel threatened by the impingement of their territory by Physics outsiders. Most of this criticism focuses on the high level of speculation in the theory, and lack of evidence: "the argument consists of merest possibility piled upon merest possibility teetering upon a teetery foundation of 'might-be-for-all-we-know's ... we judge it to be completely unconvincing and probably false." (Grush & Churchland, 1995). However, the whole point of a hypothesis is that it is speculative! A highly speculative theory is a gamble: we choose whether to risk energy on it in return for its explanatory value. And this theory promises to explain a lot, despite being high risk. *Great* scientific theories are high risk, almost by definition. So merely saying that it is speculative, without also evaluating its explanatory value, is not good science.

A second class of argument works by criticising old incarnations of the theory – specifically Penrose's initial arguments that the conscious mind is noncomputable (Penrose, 1989). However, this claim appears to be no longer central to the theory – rather it was a "tiny but extremely valuable point" which provided *inspiration* for the fuller theory's creation (Hameroff & Penrose, 1996). Grush and Churchland (1995) spent much of their paper attacking an *outdated* version of the theory, despite the fact that their criticisms had already been answered by a later version: "it would seem that from what G&C say that they have not even read, and certainly not understood, these arguments." (Penrose & Hameroff, 1995).

More mature debate over the model should involve empirical evidence and testable claims regarding the feasibility of the proposed quantum microtubule actions. For example, Spier and Thomas (1998) cite evidence that tubulins are constantly added and removed from microtubules, which would make coherence impossible. Hameroff (1998c) neatly rebuts this by citing evidence that only *non-neural* microtubules exhibit this behavior. To their credit, Grush and Churchland (1995) also raise various practical objections – the best one being that the drug *colchicine* (a cure for gout) is known to disrupt microtubule behavior but not cause anaesthesia: "this is a major monkeywrench ... surely an embarrassment". Penrose and Hameroff (1995) rebut by citing evidence that taken at the low concentrations required for gout cure, colchicine has minimal access to the brain through the blood-brain barrier; and furthermore, animal studies of higher doses *do* show cognitive effects such as learning and memory impairment. This kind of debate is good science in action – it is through facing with the best scientific attacks against it that a theory gains acceptance or is proven false (Popper, 1935).

Summary

The Penrose-Hameroff theory is high risk: it promises a very precise, low-level correlate of consciousness, placed at the fundamental level of reality – but at the expense of a mountain of speculation. It will ultimately stand or fall on the basis of empirical evidence. Given that the consciousness question is far from resolved, it is worth spending some of our energies on further investigating this theory. Whether individual scientists are prepared to take the research gamble is a matter of personal courage as well as their evaluations of the risks and rewards.

Phenomenology

What is to be explained

“The unexamined life is not worth living.”

– Socrates

We have reviewed several physical structures and processes which have been proposed as correlates of consciousness. The science used to produce these proposals relies on both high-technology experimental tools to gather data, and advanced concept structures to theorize about them. But is this science alone sufficient to find a *correlate*?

Recall that we are looking for a *transitive* correlation: a match between the *structure* of phenomenology and some structure in the brain. (Not necessarily a spatio-temporal match – it is likely that we will have to view the brain at some level of abstraction. For example, the structure might be embedded in the Fourier transform of the γ -oscillations.) The logo on the cover of this thesis shows the kind of correlation we are looking for: a matching hierarchical decomposition of each domain. Now, as we mentioned earlier, science has developed highly sophisticated tools and concepts for describing the physical domain. But there is a scandalous lack of similar tools and concepts for describing the subjective domain. On the rare occasions that science discusses the subjective domain, such as giving instructions or gathering experiential data in fMRI experiments, the language used is primitive everyday English. (More often, experimenters go to great lengths to avoid discussing the subjective at all.) This is simply not good enough. If, as some of the creature correlates suggest, the transitive correlation structure is to be found at the quantum level, then we require *equally detailed* descriptions of physical and phenomenological structures at this scale. Trying to pair up fuzzy, everyday natural language terms with high-precision physics concepts is doomed to failure. If we wish to find a detailed correlation, we must develop a detailed science for describing phenomenological structures.

Unknown to much of Cognitive Science, the development of such a phenomenological science was the *norm* in psychology until relatively recently – we are currently at the tail-end of a brief vogue of anti-phenomenological discrimination. Furthermore, *contemporary* forms of phenomenology *do* currently exist outside Anglo-American cognitivism. Rather than developing a new science of phenomenological description from scratch, can we draw upon these other traditions?

A history of phenomenologies in Psychology

Brentano (1874, paraphrased by Gardner, 1985) stressed that “one cannot conceive of thoughts and judgements, let alone study them, except by taking into account one’s phenomenological experience.” For Brentano, psychology meant the careful recollection and description of one’s inner mental life. At the turn of the twentieth century, these ideas had grown into the empirical science of *introspective psychology*, whose aim was a full description of the structure of phenomenology. Contrary to current popular opinion, introspection in its original form *did* produce useful results, which we will examine later. However, it waned for three reasons. First, a later form of introspection was flawed, and dragged classical introspection into disrepute along with it. The second reason is cultural: introspection was a predominantly German movement – and its slow, thorough project of describing a taxonomy of

phenomena was considered dull by their American counterparts: “[it] could hardly have arisen in a country whose natives could be *bored* ... They mean business not chivalry.” (James, 1890). The fall of introspection coincided with the first world war, and the German psychologists who relocated their operations in America were met with little enthusiasm. Boring (1953) describes a public demonstration of introspection at Yale in 1913 as “a dull taxonomic account of sensory events ... particularly uninteresting to the American scientific temper.” Finally, the philosophical trend of logical positivism, and its psychological sidekick, behaviorism, made all talk of internal states deeply unfashionable. Positivism also killed off the work of James (1890) – which had promised an Americanised version of introspection, concerned with the larger-scale (so less dull for Americans) forms and functions of phenomenology. Like the introspectionists, James had also made useful discoveries about the structure of phenomenology before his style went out of fashion. In sum, the reasons for the death of introspectionism were predominantly *cultural*, not due to inherent failings. Now that positivism has fallen by the wayside and Anglo-American-German relations are a little more friendly, the time may be ripe for a re-evaluation of introspectionism without prejudice.

Amazingly, almost identical movements to introspectionism have continued outside of Anglo-American academic culture throughout our ‘dark ages’ of positivism. Since Kant, *Continental* philosophy has followed an entirely separate path from our analytical philosophy, to the extent that the two disciplines are now rarely able to communicate with one other. Normally regarded as a problem, this division usefully insulated continental thought from the dogma of positivism – and allowed Husserl and others to develop a psychological philosophy of Phenomenology which flourished throughout the twentieth century. This entire body of thought seems to have been ignored by Anglo-American Cognitive Science.

There is another ‘science of phenomenology’, which predates positivism by approximately 2,500 years: Buddhism. Buddhist meditators (and their counterparts in related cultures) have evolved precision methods for examining their own internal states of consciousness which go beyond the wildest dreams of western introspections. Expert mediators may devote years of their lives to the study and refinement of introspective methods – whose ability to “detect and resolve events in consciousness” is often described as “laser-like” by westerners who become acquainted with them. (Young, 2000). While we should not necessarily accept their theoretical framework, it would be insightful to pay close attention to their methods and results. As with Husserlian Phenomenology, there may be ways of integrating them into our science. To ignore 2,500 years of thought by a culture’s finest thinkers would be pure Anglo-American arrogance.

We require a system of descriptive phenomenology in order to find the correlations of transitive consciousness, and have identified possible starting points from Introspectionism, Buddhism, and the work of James and Husserl. We will now examine each of these in detail, looking for useful *methods* and useful *results* that they have developed. We will then survey the current state of phenomenology in recent Cognitive Science, before suggesting a new approach to phenomenology which draws upon all of these traditions.

Introspection

Introspectionism was the dominant psychological movement at the turn of the twentieth century, and was pioneered by Wundt in Leipzig, and later by Titchener and Külpe. The first thing to note is that it came in two principle flavours: ‘Classical’, and ‘Systematic Experimental’ Introspectionisms. The first was good science, the second was flawed. The aim of introspection was to obtain full descriptions of phenomenological states, which were held to be complex structures built from sensory ‘atoms’ – such as points of color and frequency components of sounds. As a means to this aim, a ‘periodic table’ of atomic sensory elements and a set of rules for how they could be combined were to be discovered. Introspection was modelled on Chemistry, which was seen as the science *par excellence* of the time.

Classical Introspection

Realising that phenomenology was only accessible from the first-person, subjective view, the introspectionists rigorously trained their subjects in the arts of introspective observation and description. Subjects were typically asked to attentively observe a stimulus, and to remember the experience. Afterwards, they would recall the experience and dictate a description of it to an assistant. A single-second experience might take 20 minutes to describe in this way.

There are four obvious potential problems with the method, and the introspectionists were aware of all of them. First, it is impossible to convey an *experience* from one being to another (as famously rehearsed by Nagel, 1974). The best one can do is give a *description* of the experience. This is not a problem, since a *description* of the phenomenology is precisely what we are after.

Second, Comte (1830, cited by Boring, 1952) argued that introspection is incoherent: the mind cannot observe its own normal phenomenology as an object of that phenomenology. – because that would change the contents of the phenomenology. Instead of seeing oneself seeing red, one would see oneself seeing oneself seeing red, and that is not the normal contents of phenomenology that we are looking for. (This argument was rehearsed by Searle, 1992). However, it is not a problem because subjects are not asked to observe themselves *during* the stimulation. They are asked to *remember* the experience and recall and describe it later. Titchener's classic handbook, *Outline of Psychology* (1898) is explicit on this point: "*Be as attentive as possible ... then when the process is completed, recall the sensation by an act of memory as vividly and completely as possible.*" (his italics).

Third, there is difficulty in describing an experience in natural language – words may be ambiguous and imprecise. This is often cited as the grand failing of introspection. The introspectionists were well aware of the potential problem, which is why they spent so much effort training their subjects to use standardised language to give more precise descriptions. Titchener (1898) states "The terms chosen to describe the experience must be definite, sharp and concrete ... words are little blocks of stone, to be used in a mosaic". Titchener did not allow subjects' data to be published until they had practiced 10,000 trial descriptions in the standard language. A genuine problem, however, was the practical one of standardising this language *between* laboratories – each lab had its own pool of trained subjects, but the labs inevitably picked up their own dialects and habits. For example, Boring (1942) chronicles the famous debate between the Titchener and Külpe schools' classifications of atomic sensations: with Titchener claiming "more than 44,435" elements compared with Külpe's "fewer than 12,000". This was due to a lack of standardisation between the two teams. But *within* each laboratory, there were no such problems. All that was needed was a global standard.

Finally, there is the problem of conceptualisation. Classical Introspection wanted a phenomenological description couched in purely terms of sensory atoms. But we often do not perceive objects just as low level sensory patterns – we use past experience and concepts to group and categorise stimuli. So instead of reporting sensations of roundness, redness and hunger, we might report seeing an apple. Different subjects have different pasts and different conceptual frameworks, so might report different phenomenologies for the same stimulus. Titchener was the most extreme of the introspectionists with regard to this problem: he insisted that all meaning and inference be kept out of descriptions, since they were deduced from, and were not part of, the stimulus. When they crept into description he discarded them as 'stimulus error'. "When we introspect, we must be absolutely *impartial* and unprejudiced. We must not let ourselves be biased by any preconceived idea ... we ought to take the facts precisely as they are." (Titchener, 1898.)

Contrary to current orthodoxy, Classical Introspection did produce useful results. For example, Hering (cited by Mangan, 1993) introspected his perception of color – noting facts such as that yellow is perceived as a single sensory element and not as a mixture of red and green. From this he was able to deduce the structure of the underlying mechanisms. These deductions were much later shown to be correct by De Valois and De Valois (1975, cited by Mangan, 1993). This is just the kind of observation we need when looking for transitive correlates of consciousness.

Classical Introspection, like our quest for correlates, aimed to provide full descriptions of the elements and structures of phenomenology. Aside from its failure to establish an inter-laboratory language for reporting results, its methodology was not *inherently* flawed. It aimed to describe the *low-level*¹ phenomenological world *as experienced by the subject* – in which case the subject's memory could usually be relied upon to give an accurate description of *apparent* events. Averages over many subjects were taken to reduce individual errors. (Titchener, 1898). *Practical* problems arose from the use of (standardised) *natural language* for reporting: First, this was time consuming so memories may have faded and become confabulated towards the end of the reporting process. Second, completely standardising natural language between laboratories appears impossible due to language's complexity and the appearance of local dialects. Perhaps these *practical* problems could be resolved, and

¹ I use 'low-level' rather than 'preconceptual' because I will later argue that even the lowest-level experiences are conceptual.

introspection revived, by utilising modern information technology systems to replace natural language as the reporting media?

Systematic Experimental Introspection

As mentioned earlier, *Classical* Introspection was brought into disrepute by the activities of its rebellious cousin, *Systematic Experimental* Introspection (hence, SEI). This school was founded by Külpe, who left the Leipzig school following a disagreement with Wundt's classicalism. Külpe believed that introspection should not be confined to the Classical project of describing the passive perception of stimuli. He wanted to extend the project to cover *active* thought processes, such as problem solving. He thought that the mechanisms of thought were discoverable by introspection.

The first wave of SEI, around 1901-1905, used the classical introspective framework of reporting sensory elements and structures, but now applied to active cognition. The then-remarkable conclusion was that the phenomenology of problem solving consisted only of a series of images of various stages of the solution process, with apparently "no hint as to how or why they were formed". (Reviewed by Boring, 1953.²) There are two possible conclusions that can be drawn from this: First, that active cognition is essentially an unconscious process, and we are only conscious of the (intermediate) results of it. Or secondly, that active cognition uses a new class of phenomenological entities which are not revealed by Classical Introspection. Today, most would draw the first conclusion. For example, the Penrose-Hameroff theory postulates a cycle of unconscious processing followed by moments of consciousness; and Baars' Global Workspace contains many unconscious agents, only occasionally stepping into consciousness when they need to share intermediate results. (Interestingly, Freud was coming to similar conclusions in the period contemporary with Külpe.)

The SEI school drew the opposite conclusion: that there are elements of phenomenology which are not sensory, and hence are not reported by the Classical Introspective method. Hence the second wave of SEI trained its subjects to report their phenomenology in terms of both sensations and these new elusive 'acts'. Their existence, even at the time, was controversial. The Classical Introspectionists denied them – this was the famous 'imageless thought' debate between Titchener and Külpe. Külpe explained acts' elusiveness by claiming that they are not noticeable *during* cognition, only by introspecting afterwards. His critics countered that this was a licence for subjects to confabulate stories about how they solved the problems unconsciously.

The 'imageless thought' debate is often used to denounce the whole of introspection as incoherent. For example, Güzeldere (1997) caricatures it as "a stalemate of, 'You cannot experience X,' of Titchenerians versus 'Yes, we can!' of Külpeians". However, this denouncement is unfair. Güzeldere portrays the debate as an unverifiable clash of empirical results – one school reporting discoveries of X (imageless thought) and the other reporting its absence. But the two schools were using *different theoretic frameworks and experimental methods*, one of which (supposedly) observed acts and the other of which did not. The clash was not over empirical results, but was the question of whether to allow acts into the ontology. That is, the Titchener school assumed the first, 'unconscious processing' conclusion from the first-wave experiments, and the Külpe school assumed the second, 'imageless thought' conclusion. We now have good reason to believe that the 'unconscious processing' conclusion is the right one. So the Külpe assumption is wrong, acts do not exist, and the reports of Külpe's second-wave subjects were confabulations, not true descriptions of the (unconscious) mechanisms of thought. So we could say that Titchener was correct in the imageless thought debate and Külpe was wrong. The imageless thought debate was about frameworks, not empirical results, and its existence does not demonstrate the incoherence of Introspection *as a whole*. However, our modern knowledge of unconscious processing *does* denounce the validity of Second-Wave Systematic Experimental Introspection. (It is a sign of contemporary prejudice against introspection that Güzeldere's does not even show awareness of the distinctions between the different introspective frameworks.)

So second-wave SEI is debunked. However, the very fact that we can pose the imageless thought question depends on a *result* from it's first wave: the sound empirical finding that phenomenology during thought consists of a series of intermediate images. (Recall that this was found using Classical

² Most of the introspectionists published only in German, hence the use of secondary sources here.

Introspection methods of phenomenological description). This is in itself a useful result – as it provides us with phenomenological data for which we could find physical correlates during thought.

Jamesian Introspection

Contemporary with the atomic-structural German introspection, William James (1890), an American, was interested in the higher-level structures and functions of phenomenology. His introspection took a looser, fuzzier, top-down approach. Importantly, he noted that our everyday phenomenology does not appear to consist of the structures of *low-level* sensations observed by the Germans' subjects: these sensations were only present in their phenomenologies because they had been *explicitly instructed* to pay attention to them. Normal phenomenology for James would consist of 'an apple' rather than the artificially-induced 'a sensation of redness and roundness'. James aimed to "begin with the most concrete facts ... [this method] will discover in due time the elementary parts, without dander of precipitate assumption." He was also keen to follow neuroscientific findings, as a means of guiding his phenomenological investigations. (It is interesting to note the symmetry of this approach with the German method of using phenomenology to inspire neuroscience.)

The I and the Me

One of James' key distinctions is between the subjective, immediate experiencer - the 'I' - and that experiencer's concept of its physical and historical *self* – the 'Me'. The contents of the experiencer's phenomenology contains many *objects*, such as apples, and the self is just one of these objects. The immediate I adopts the remembered history of Mes, and the I's basic function is to preserve the existence and contentment of the Me. This distinction suggests that the I can possibly exist without containing a Me. This is important for modern theory, since it suggests that so-called *self-consciousness* (i.e. consciousness whose phenomenology contains the self-object, the Me) is just a special case of consciousness, rather than an important necessary or sufficient condition for it as some contemporary theorists have suggested.

(An interesting parallel to the I/Me distinction is found in Wittgensteinian thought (e.g. Anscombe, 1960), which notes that it is possible to misidentify the Me but not the I.)

Changing sensations versus constant objects

James disputes the German assumption (and its history, going back to Locke) that phenomenology is structured from a set of atomic sensations. Instead, our perception of the same stimulus is altered by the context in which it is presented, and by our ever-accumulating history of experience. (The history argument is inspired by neuroscience: though predating Hebb's formalisation, James knew that every perception could modify the brain in some way.) Low-level sensations have little importance in Jamesian phenomenology; rather than getting the same sensation twice, "*what is gotten is the same OBJECT ... the sameness of things is what we are concerned to ascertain.*" James implies that unless we are instructed to pay specific attention to low level sensations (as in the German experiments), we may not even be conscious of them – only of the whole objects that they evoke. Interestingly, a further cause of differing sensations of the same stimuli is *mood*. For example, if mood becomes more depressed, then "what was bright and exciting becomes weary, flat and unprofitable". While James does not develop a detailed phenomenology of moods, their presence and effects are important points to note.

The stream of consciousness

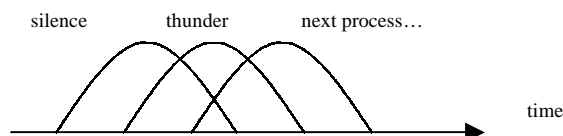
James was especially interested in the *dynamics* of phenomenology, described by his famous 'stream of consciousness'. Like the first-wave systematic experimental introspectionists, he observed that phenomenology appears as a series of images, with little indication of how or why each image is produced. He calls the images and transitions the *substantive* and *transitive*³ states, respectively; and likens their cycle "a bird's life ... an alternation of flights and perchings":



³ Not to be confused with the 'transitive' of 'transitive correlate of consciousness'.

He suggests that the rhythm of language reflects the same pattern: with an image formed at the end of each sentence, but with little in consciousness during its transitive reading. His answer to the 'imageless thought' question sides with Wundt (and the modern view): there *is* much more to thought than phenomenological images, but it happens in the almost unobservable transitive periods. We sometimes get a glimpse of the edge of transitivity, but by its very nature, if we try to focus upon it, then it stops being transitive! Trying to form a mental *image* of the elusive transitivity is like "trying to turn up the gas ... to see how the darkness looks."

Although phenomenology is a series of *discrete* substantive states, the difference between successive states is always small; there are no *abrupt* changes. Even if one state is the hearing of silence and the next of thunder, the change is gradual. James explains this apparent paradox: we aren't just aware of 'thunder' but of 'thunder-breaking-on-silence'. There are multiple processes active *simultaneously* – the silence process gradually fades and the thunder process gradually becomes prominent:



The fringe

Apart from the 'stream', James' other great contribution to phenomenology is his notion of consciousness containing a *focus* and a *fringe*. The vivid, central, introspectable parts of phenomenology are surrounded by a penumbra of vague, semi-conscious information. This fringe seems to be concerned with context and *rightness*. (e.g. we receive a shock if we read "plumber's bill" in the middle of a thesis on consciousness, because the word is out of context; it is not 'right'.)

The contents of the fringe are concerned with *associations*. If our phenomenology is focussed on an apple, then the fringe presents us with relevant background knowledge of apples, from our past memories. This helps us to conceive the stimulus *as* an apple (rather than just redness and roundness) and gives us a sense of context. We get a vague 'feeling of knowing' about this related information – such as information about pears – telling us that pear information is available if we would like it. We don't *know* information in the fringe, but we know that we know it. James says true knowledge of an object is knowing its relations (i.e. having them present in the fringe) – *acquaintance* is merely having the object in focus (with an empty fringe).

Like transitive states, the contents of the fringe usually elude introspective efforts. Trying to focus on the pear in the fringe actually recalls the pear from memory and places it in the focus. James was not acquainted with eyetrackers, but an experiment recounted by Dennett (1991) provides a good analogy: the subject looks at a blurry screen of text, but the screen is eyetracked so as to replace the blur in the subject's small fovea with a crisp version of the text. The subject has been told that the screen is blurry, but every time he moves his eyes to look at a blur, it is replaced by crispness. In the same way, we are told by James that the fringe is blurry, but every time we try to move our attention there, the blur is replaced by pure knowledge.

How, then, did James deduce the structure of the fringe? In an ingenious sequence of introspections, he managed to set up a condition in which the fringe contents was not immediately replaced by its 'full-content' knowledge. The condition is known to most people as 'tip-of-the-tongue' or 'feeling-of-knowing': when asked to recall an obscure piece of information, such as 'What is the capital of Brazil?', we are aware that we know the answer – that much is presented to us in our fringe – but we have trouble actually recalling the full content. This feeling may last for several seconds – so can be studied introspectively.

In sum, James' key results are that phenomenology usually consists of high-level *objects* rather than sensations; the self is just one of these objects and is not necessary for consciousness itself; dynamically, phenomenology is a series of discrete substantive states, produced by unconscious transitive processes; each substantive state contains a fringe bearing semi-conscious contextual associations. We could perhaps redraw the earlier state diagram to show the fringes:



Husserlian Phenomenology

Roughly contemporary with the introspective efforts we have seen so far, Edmund Husserl (e.g. Husserl 1972; reviewed by Sokolowski, 2000) worked within the Continental tradition to produce a very similar philosophy, which he named ‘Phenomenology’. (There is a danger of confusion here, as his term is now used (as in this thesis) to refer to *any* method for the first-person study of consciousness – but when he and his followers use the word, they mean only the Husserlian style. Here we will use ‘Husserlian Phenomenology’ to avoid confusion.) Husserl was more a philosopher than a psychologist, and his Phenomenology is a grand philosophical system which concerns itself with much larger domains than our present concern of describing the contents of consciousness. We will restrain ourselves to the latter here. But we should note that his objective is not primarily to establish a rigorous *practice* for phenomenology – more a way of thinking.

Husserl is a difficult philosopher to read, not least because of the difficulty of translating his German, but also due to his larger project. Interpretation is not made easier by his ever-growing caravan of followers, commentators, hermeneuticists, postmodernists, deconstructionists and other hangers-on so beloved of the Continental tradition. To the Anglo-American reader, it is quite a shock to see that Husserlian Phenomenology journals are not filled with empirical phenomenology results, but with commentaries-on-commentaries on Husserl’s philosophy. In order to avoid becoming part of the aforementioned caravan, we will be brief, and examine only the ideas which can be borrowed for our current project.

Central to Husserlian Phenomenology is the concept of *intentionality*: A conscious representation (the *noesis*) is *about* an object in the world⁴ (the *noema*). This is a rebellion against the Cartesian-Lockean tradition, in which we are primarily aware of our own ideas, not the world. For Husserl, objects in the world are primary. Further, we can intend objects in different ways: for example, by direct sensation, remembering and imagining. Husserl’s aim for psychology is “to investigate systematically the elementary intentionalities ... processes [and] structural composition” (Husserl, 1972). This sounds like what we need in our search for consciousness correlates.

Husserl’s method for phenomenology is called *reduction*⁵. The essential problem with intentional objects is that they are *transparent*: instead of being aware of them, we see through them to the world-objects that they represent. This usual mode of perception Husserl calls the *natural attitude*. In reduction, we deliberately prevent the natural attitude. We strive to look and describe at our experiences *themselves*. This is done using the *epoché* (literally, ‘stopping’): the constant paying of attention to the stopping of conceptual thoughts. We strive to focus on the redness and roundness in our experience, rather than the apple. If a concept comes, we *bracket* it (as in ‘put it into brackets’) – that is, we look down upon the concept as another frozen part of the phenomenology. Rather than seeing the apple, we now see our phenomena of redness and roundness, and also see the apple *noesis*. By ‘stepping out from ourselves’ in this way, we get to look *at* our intentional objects rather than through them. Husserl calls this outlook the *phenomenological attitude*.

Reduction is similar to Classical Introspection in that it aims to give a low-level, preconceptual account of ‘raw’, sensory phenomenology. It differs in two main ways. First, this account is the *only* goal of Classical Introspection; for Husserlian Phenomenology it is just the beginning – by bracketing, we later examine more complex intentionality structures. Second, reduction is performed at the time of perception, as opposed to Classical Introspection’s retrospective recall.

Husserlian Phenomenology is less taxonomic than Classical Introspection in its reporting style. Like Jamesian introspection it aims to give a ‘broad-brush’ picture of the different types of high-level elements and their interactions, rather than Classical Introspection’s precise low-level elemental

⁴ Or possibly to a fictitious object which is not in the world.

⁵ Not to be confused with contemporary ‘reductionism’ in mind-brain identity theories.

descriptions of momentary phenomenology. We will now quickly survey some of these high-level findings of Husserlian Phenomenology.

Once we have identified the low-level preconceptual intentionalities, we can catch and bracket higher-level concepts which are based upon them. We find that we can bracket *multiple layers* of intentionalities upon the same low-level perceptions. Hut (1999) gives a simple concrete example. Holding a pen at arm's length, he notes his low-level intentions of the colors and shapes of his immediate vision. Moving the pen towards and away from him, he notes that the shapes at this level get larger and smaller due to perspective. Next, he brackets his higher-level intention of the whole pen in space. This pen does not change size during the motion. *He is simultaneously aware of the pen at two different levels of abstraction.* Further, Hut is a physicist, and so can form a third intention of the pen: intending it as a collection of atoms. (This is a more complex kind of intentionality, since the intended object, the set of atoms, is not immediately apparent to sensation.) We could imagine other people entertaining even more intentionalities: the pen-maker may intend the pen as a collection of its mechanical components (ball, ink, lid etc.); the poet may intend it as a creativity device.

More complex simultaneous intentionalities occur when we intend *symbols*. Seeing a painting, we can intend the paint and canvas, but also the objects which it depicts. Similarly for visual and audio presentations of words. If watch an 'innovative' interpretation of *A Midsummer Nights Dream*, say one performed without any words and set in a 1970s nightclub⁶, we intend both the pure play itself *and* the particular interpretation. In these cases, as with the atoms of the pen, we are capable of intending objects which are not actually present in our immediate sensation. Husserl calls this *empty intention* (as opposed to *intuition* of immediate sensation).

Note that the multiple layers of Husserlian Phenomenology provide a resolution of Classical and Jamesian introspections. Classical Introspection assumes that phenomenology is made of low-level elements; James claims it is principally only higher level concepts. Husserl says that either or both are possible. We can choose to concentrate on the low or high levels, but more usually entertain several intermediate layers at the same time.

Our sensations are constantly changing, yet our high-level intentionalities can remain static. How is this possible? Husserl developed the key notion of *identity* (of the high level) in a *manifold* (of a lower level) to investigate and describe this process. A manifold is the set of all low-level intentions which are sufficient to bring about a high-level intention. Sokolowski (2000) gives an example of this process occurring between multiple levels. Suppose we are looking at a cube. We intend the cube as a whole. But we never see the whole: only three sides at a time. The visible sides are a cue for the whole cube – they are part of its manifold. Once we grasp the cube's identity, we can then emptily intend the far sides as a lower layer of the whole. Similarly, if we are walking around the cube, and so are faced with constantly varying perspectives of the sides, how to we intend a the identity of a side given that our low-level perception of it is changing? Again, they are part of its manifold. Finally, even if we have a constant perspective on a side, how to we maintain its identity through time, from moment to moment? This is a *temporal* manifold.

Husserlian Phenomenology does not attempt to explain *how* manifolds are formed – it aims to *describe* them. Like the other first-person methods we have seen, Husserlian Phenomenology is descriptive. Explaining the underlying non-phenomenological mechanisms is the task of third-person science. Despite its initially fierce terminology, vague empirical methodology, and hermeneutical hangers-on, it adds useful illumination to our project of describing the structure of consciousness – most notably through its concepts of *multiple* and *empty* intentionalities. The former resolves the Classical/Jamesian levels dichotomy; and the latter is reassuringly reminiscent of James' fringe.

Buddhist Phenomenology

Unlike western academic philosophy, Buddhist thought is traditionally bound to practice: rather than just talking about phenomenology, Buddhists 'philosophers' actually *do* it. Their philosophy, psychology, arts and ethics are linked by the activity of *meditation*. There are countless different schools and traditions of Buddhism and its relations, which we do not have time to survey here. Neither can we examine the grand philosophical schemes of which Buddhist phenomenology is only a

⁶ Such things only happen in Edinburgh. (This particular production was called *The Donkey Show*.)

small part. Instead, we will try to draw out some of their phenomenological methods and findings which could be useful to our present thesis.

In stark contrast to western third-person science, Buddhism begins with the first-person: “All phenomena are preceded by the mind. When the mind is comprehended, all phenomena are comprehended.” (Buddha, trans. Wallace, 1999). Understanding the contents and processes of the conscious mind is thus the primary focus of its science. Initial assumptions are that our everyday phenomenology only shows the surface of a deep underlying structure of consciousness, and that it is possible to learn to introspect this structure through the highly skilled process of meditation. The ‘disciplined mind’ is a highly-tuned, high-precision tool for scientific investigation. As one would not use the naked eye for studying quasars, so one would not use the untrained, ‘dysfunctional’ mind for studying psychology.

Wallace (2001) distinguishes two disciplines within meditation practice. *Quiescence* is the skill of gaining access to the structures; *Insight* is the science of exploring and theorizing about them. Quiescence is achieved by cultivating *mindfulness* and *introspection*.

Mindfulness is the capability of prolonged, perfectly focused attention. Western Psychology commonly holds that ‘spotlight of attention’ is constantly in motion, rapidly scanning the wide world with its narrow focus. (e.g. this is assumed by Crick and Koch, above.) Nineteenth-century research suggested that attention cannot be held in place for more than about 3s (Wallace, 1999), and James’ introspection (1890) concluded that “No-one can possibly attend continuously to an object that does not change.” However, through careful training, often taking several months or years to complete, meditators learn to defy this limitation and to hold their attention in place by conscious effort. Complementary to mindfulness is *introspection* (not to be confused with the western meaning of the word). When mindful of an object, it is easy to ‘drift off’ into a state of lower wakefulness. Introspection is the skill of *monitoring* and maintaining the level of high alertness.

Wallace (1999) details a typical training process: The student first cultivates mindfulness through a progressive series of attention exercises, before learning the skill of introspection. As with most skills, mindfulness and introspection gradually become automatic, and by the end of training the student is able to perform both *without conscious effort*. This is known as *balanced placement*. Now that quiescence has been learnt, insight can begin.

Layers of consciousness

Exploration of consciousness is performed in a series of stages, in which consciousness is found to consist of a number of layers. The number and order of the layers differs between traditions, but the general process is the same: each stage of exploration consists of ‘peeling off’ a layer to reveal the more fundamental one beneath it. To understand a layer is to transcend it. Once all the layers are removed, the meditator is left to experience consciousness in its purest, uncluttered form. (e.g. Shear, 1996).

To begin the process, attention is focussed on an object. Sometimes this is physical object - such as a rock or religious statue – sometimes it is abstract, such as a verbal *mantra* (a repeated slogan used in westernised Transcendental Meditation) or *kōan* (a paradoxical parable used in Zen Buddhism). The purpose of this is to achieve balanced placement of attention. Once balanced placement is achieved, attention can then be transferred inward to a higher layer of consciousness. (Such layers are difficult to focus on, and perception of them is lost if attention falters – hence the need to cultivate balanced placement before looking for them).

Reminiscent of Husserl’s ‘bracketing’ and Classical Introspection, a common Buddhist layer is that of sense-images. Attention can be focussed on the mental images of sensation, rather than seeing through them to their intentional objects. Husserl’s famous dictum, “Back to the things themselves” is mirrored by Buddha: “In what is seen there should only be the seen; in what is heard, only the heard.” (Buddha, trans. Wallace, 2001). (Buddhists distinguish six, rather than five, senses – with mental perception (e.g. imagery) being the extra one.)

A further layer, above raw sensation, is that of conceptualised objects. Paul (1981) details Yogaacarin Buddhist claims that this layer is constructed with reference to the *self*: The self-object is at the centre

of the conceptual phenomenology, and together with sense-images, its past experiences shape the formulation of the conceptualised objects surrounding it.

Pure consciousness

In the ultimate layer, we attend to pure consciousness itself, transcending not only sense-data, but also concepts *and even the self*. Consciousness can “transcend its own activity” (Shear & Jevning, 1999). It ceases to be ‘transitive’ consciousness-of and becomes pure ‘creature’ consciousness. This provides an important counterexample to well-entrenched western assumptions. Hume’s bundle theory (1748) suggested that consciousness is nothing but a bundle of perceptions – Buddhism shows that there is something else independent of them. The transcendence of the self is almost unique: even though trying to doubt everything, Descartes still assumed that ‘I am a thinking thing’. Similarly, Husserl tried to bracket every concept, but still assumed the existence of a ‘transcendental ego’. In the west, only James’ I/Me distinction seems to have come close to the idea. In pure consciousness, there is no ego, no self, there is only ‘I am’. Empty phenomenology can exist; consciousness without intentionality is possible.

Those who have experienced pure consciousness typically try to explain it as a feeling of pure joy, bliss, happiness. (Though obviously the experience cannot be conveyed by words; only described by them.) How could this be, if they have risen above the layer of feelings and emotions? Perhaps the answer is that consciousness in its pure state *is* the feeling of bliss; and the additional *contents* of consciousness act as filters of its light. Different phenomenal objects filter out different parts of this feeling to produce different emotions. As we go through life, we learn new, more complex concepts – ever more intricate ways of obscuring the light. Perhaps this is why to be newborn and innocent of such structures is to be truly happy? As we grow old and experienced, we cloud pure consciousness with our concepts (especially our self-concept). The study of ‘feels’ of phenomenal objects would be a fascinating future research area, but is one which does not concern the present thesis. We are currently only concerned with the *presence* of the contents of consciousness. We are not yet trying to explain why certain phenomena feel the way they do⁷.

Moments

We have so far looked at the *structure* of consciousness, but Buddhist phenomenology has also examined its temporality. Hameroff and Penrose (1996) provide a review of temporality findings. Expert meditators have refined their awareness of time, and report that phenomenology appears to be quantised into discrete events. Buddhist texts portray each event as instantaneous, and some schools even report numerical observations of their frequency; the reports range from one event every 0.13ms to every 20ms. This idea of discrete *moments* is what James reached with his notion of Substantive states – but the precision ‘contemplative technology’ (as Wallace is fond of calling it) of Buddhism has provided us with a more accurate description than James’ introspection. Note that the 20ms result equates to a moment frequency of 50Hz, which falls within Crick and Koch’s γ -oscillation range.

The tradition of Buddhist meditation seems to have much to offer us in our search for a descriptive science of phenomenology. It has developed techniques for refining awareness and for introspecting which are more complex than the crude methods used by the Introspectionists. While its strategy is similar to the thinking of Husserl, it has been developed into rigorous practice. It can shown us that pure ‘creature’ consciousness without content is possible, and has augmented James’ ‘flights and perchings’ with quantitative measures of moment frequency.

Whatever happened to Phenomenology?

We have reviewed phenomenological methods and results from traditions outside the twentieth century Anglo-American behaviorist/cognitivist tradition. We will now survey the recent and current states of

⁷ I would however like to suggest a brief suggestion for the evolutionary purpose of phenomenal pleasure and pain: Artificial life forms (e.g. Fox, 1999) have been designed to function without consciousness, and typically employ ‘punishment’ and ‘reward’ systems for guiding their behaviors. A major problem with such creatures is that they can accidentally evolve to mutate these systems, e.g. changing their a part of their program from ‘be in pleasure when you see food’ to ‘be in pain when you see food’. This is clearly bad for survival. If, alternatively, the pleasure is an *inherent* property of the phenomenal food object, by virtue of that object’s filtering of pure consciousness, then this kind of mutation would not be possible. We are not able to become happy simply by changing our definition of happiness from what we need to something more obtainable!

phenomenology *within* the latter tradition – and will see the potential beginnings of a renaissance after the behaviorist dark ages.

The Dark Ages

It is informative to examine what happened to the fruitful neuroscience-guided-by-introspection programme typified by James. Essentially, the positivist movement outlawed first-person experience as a source of data, claiming it to be unverifiable and therefore meaningless. Scientifically, consciousness was not allowed to exist. In its most dogmatic psychological form – behaviorism – the positivist view prohibited *all* non-observable entities from theory: only immediate stimulus-response relationships were allowed.

Positivism was later shown to be untenable, notably due to Quine's (1953) demonstration that meaning inescapably involves unverifiable linguistic assumptions in addition to any methods of verification. Together with the rise of Turing Machine and Finite State Automata, this opened the gates of psychology to cognitivism. In addition to finding immediate stimulus-response relations, internal *state* was postulated in the mind to allow explanation to more complex behavior. But despite the allowance of these computational states, cognitivism still implicitly holds on to the behaviorist consciousness ban. These states have nothing to do with consciousness: they are supposed to be posited – as the physicist posits his quarks – in order to give the most parsimonious mechanism for explaining observed, objective data.

But this parsimony doctrine is not followed in practice. Although discussion of consciousness is banned from the literature, we find that researchers *are* still using their own phenomenological observations to guide their models. Though they would never admit to it in publications, most cognitive modellers in areas such as attention and working memory are postulating states which resemble their phenomenology rather than those which are strictly most parsimonious.

This ludicrous situation reminds me of research in the foundations of mathematics. Since the beginning of civilisation, humans have been using a system of 'folk' mathematics – learning useful rules of arithmetic and geometry. *Later*, researchers (such as Russell and Whitehead) have tried to formulate the foundations upon which these folk systems are supposedly based. Now, many mathematicians hold that mathematics exists in some Platonic sense, independent of human use; its foundations, even before we have formulated them, should already exist to be discovered. But in fact there are infinitely many systems of axioms and rules, each of which would give a different mathematics. Foundation researchers have not discovered the one-and-only 'pure' foundation; they have merely looked for a system which 'conveniently' happens to produce the same large scale systems that folk psychology has already been using. In fact they went to rather extreme lengths to produce a convoluted foundation in order to escape apparent paradoxes in folk mathematics⁸. From an infinity of equally 'pure', 'perfect' systems, they pick out the one which corresponds to the essentially arbitrary system they already happen to use; then go on to claim that this is the only pure Platonic system.

In the same way, psychologists are faced with a multitude of possible complex systems of internal states to postulate to explain behavior. But, like folk mathematics, they have a secret prejudice that they would like to justify – the phenomenology that they already experience. So they choose the system which matches it best: their theoretic postulated states just 'happen' to match their phenomenology. But this is never reported – instead there are just vague confabulations about 'most parsimonious model'. Like voting Tory, many people are *doing* phenomenology but few will publicly admit to it.

A new renaissance

It is only in the last few years that phenomenology-driven cognitivists have begun to 'come out' and call for a new public acceptance of phenomenology. The call is for phenomenology, psychology and

⁸ Folk mathematics defines numbers by example: three is explained by showing the student examples of three things. In order to avoid Russell's paradox, numbers instead have to be defined in a vastly more complex terms of sets – concepts which are completely alien to users of folk mathematics. Perhaps mathematicians should just accept that folk mathematics is not perfect, but is an evolved collection of practical, useful, set of rules for making predictions about the world. Most users of folk mathematics don't mind it containing occasional paradoxes as long as it gives them good everyday predictions. Like language, there *are* no inherent foundations. Foundations are *invented*, but not *discovered*.

neuroscience to work together to discover the structures of both phenomenology and the physical brain. By using each others' findings to provide 'mutual constraints', a 'circulation' of 'mutual enlightenment' will ensue. (Verala's term for this new science is 'neurophenomenology', in deliberate opposition to the Churchlands' materialist 'neurophilosophy'.) Essentially, this is just a return to what James and the Introspectionists were trying to achieve: the use of first and third person investigation to guide each other. Notable protagonists include Shear, Velmans, Revonsuo, Flanagan and Varela; and the last three years have seen two special issues of the *Journal of Consciousness Studies* (Vol. 6, Nos. 2-3 (1999); Vol. 8, Nos. 5-7(2001)) devoted entirely to their discussion of re-integrating phenomenology with Cognitive Science. We will here examine some recent suggestions and findings.

Baars: Psychology as Phenomenology

We should start with Baars' claim that "there is already a field of systematic phenomenology, it's called psychology" (Baars, 1999). He points out that psychology generally relies on verbal instructions and reports, and that reportability is generally taken as an index of consciousness. Cognitivist investigations into the structure of working memory and awareness can now be reinterpreted as descriptions of the structure of consciousness. However, he admits that psychology has only so far described the 'surface structures' of phenomenology - and has not yet examined the layers or the fringe found by the phenomenologists we have reviewed. The point is that our search for the structure of phenomenology need not start from scratch - we can build on what we already know from third-person science.

Second person methods

A key difficulty with phenomenology is its lack of intersubjectivity. It is essentially a first-person experience which cannot be conveyed exactly to others. This conflicts with one of science's most basic principles - the need for public reports. The traditions we have examined all depend on a process of *description* of the experience in language, but how should we reconstruct the phenomenology from the report, given the vagueness of language? This question has been the subject of much recent debate.

So-called *second-person* methods for interpreting the reports attempt to approximate a *reconstruction* of the subject's experience in the phenomenal world of the report reader. (e.g. Naudin *et al.*, 1999; Braddock, in press). This is to be done by the reader familiarising himself with the subject's culture, history and language, and doing his best to imagine 'what it would be like'. Naudin *et al.* have applied such a method with Husserlian reduction as a technique in psychiatric practice. For example, the traditional method for diagnosing schizophrenics is by making observations of the *types of delusions* they report - and how those delusions differ from the 'correct' perception of the world. Naudin *et al.* instead use the second-person technique. They treat the reports as descriptions of *real experiences*, and imagine what it would be like to have them - the instinctive sense of 'delusion' is bracketed. From these reconstructed experiences, they look for common themes in their semantics - for example, many of them concern bodily invasion. And it is this characteristic - body invasion - that is the indicator of schizophrenia. The types of delusion are just surface consequences; phenomenology reaches the deeper symptom.

Dennett (1991) thinks that intersubjective familiarisation is unworkable - we can never come to understand words in the same way as another well enough to reconstruct their phenomenology in our own experience. Instead, he proposes *heterophenomenology*. For Dennett, the subject's report is to be treated strictly in the third person by the scientist. Taking the 'intentional stance', the scientist is to *theorize* about what possible phenomenology could have caused the subject to produce the report. Rather than reconstructing the phenomenology in his own experience, the scientist treats the whole problem in the third person. He compares this skill of heterophenomenological interpretation to that of interpreting historical texts - since a knowledge of the culture and history surrounding the report is needed to understand how it could have come about⁹. However, it is difficult to see how the understanding required here is any less than that required for second-person re-experiencing. The only obvious cases where heterophenomenology beats second-person are rare Nagelian situations in which elements of the subject's experience are *inaccessible* to the scientist, for example if the subject is synaesthetic¹⁰, or if the scientist is blind. (c.f. Nagel, 1974.)

⁹ Perhaps the Husserlian commentary entourage would be better reemployed in the service of heterophenomenology?

¹⁰ Synaesthesia is a rare condition in which senses 'blend into one another', e.g. a trombone may sound red to a synaesthetic.

The problem of imprecise word meanings has been attacked in a different way by Marbach (1993, reviewed by Gallagher, 1997). Marbach is inspired by the Classical Introspectionists' attempt to create a standardised version of natural language for reporting, but takes the idea further and creates an entirely new symbolic language for the purpose. For example, the perception of object X is denoted (PER)X, the replaying of perception is denoted [PER]X. Memory is defined as "a perceiving of X bestowed with the belief of it actually having occurred", denoted (REP P * [PER] X). (Examples taken from Gallagher, 1997). This allows Husserlian intentionalities of perception, re-perception, imagination and so on to be clarified more easily than in natural language. Note that Marbach's language says nothing about the structure *within* each intentionality, such as how the perception is comprised of elements; it is just a method for distinguishing between the different kinds of intentionality. Could we build on the idea and extend it to provide full phenomenological descriptive power? This is all beginning to smell suspiciously Carnapian: a formally defined perfect language. We should ask how the terms would get their meanings, such that they would be intersubjectively verifiable. On reflection, it seems that the only way we could teach subjects the meanings of terms would be by presenting them with a series of *exemplars* of their usage. But surely the way in which the subject generalised the term-concept would depend on their past experiences? For example, imagine teaching subjects a formal term CAT by showing them a set of three pictures of cats. One subject is already familiar with many kinds of animals, so 'correctly' learns that CAT refers only to *cats*. But another subject is not familiar with animals – he has never seen dogs, horses, sheep etc. For him, the notable features of the exemplars are not the specifically cat-like features, but the more general animal features such as having four legs. So he 'incorrectly' generalises that CAT refers to *animals*. Marbach has given us a useful idea – that of producing a standardised representation for phenomenology reports – but it seems that a *symbolic* language is doomed to failure due to problems with defining the symbols. This brings us neatly to the work of Revonsuo.

Revonsuo: dreams, concepts, and virtual reality

Revonsuo (e.g. 1995, 2000) is interested in the distinction between conscious (phenomenological) and unconscious (non-phenomenological) perception in the brain. We have seen so far that phenomenal worlds are comprised of a series of layers, ranging from the low-level elements observed by the introspectionists, to the high-level concepts of Husserl and the Buddhists. Revonsuo is concerned with which of these layers are necessary for phenomenal experience.

During dreaming, the low-level visual areas of the brain are not active, but the high-level concept systems are. This fits with our experience that dreams appear vague, lacking detail¹¹. We may be aware of a tree in our dream, but are unlikely to be able to count how many branches it has. It just appears as a TREE concept. Revonsuo's main concern is to 'carve off' non-essential areas of the brain from the search for correlates of consciousness – so he concludes that we need not look in these low level areas. Our concern, however, is the description of phenomenology. We can interpret this work as showing that *high-layer concepts can exist in phenomenology without the low-level sensory elements* that the introspectionists looked for.

As the Buddhists progressively strip away layers of phenomenology, so Revonsuo uses dream-consciousness to strip away the sensory data layer. Dream reports provide him with access to a purer form of consciousness; some of its non-necessary features have been disentangled and separated off. Having obtained these 'pure sample', he asks the question: 'What is common to all of them? What now seem to be the essential features?'. The answer: a *self* in a *world*. Revonsuo notes that almost all dream reports begin with a description such as "I am standing in a park, by a tree, and I am in a hurry". The park and the tree are the world; the 'I' and the knowledge that 'I am in a hurry' show the presence of a self-concept located within the world. Revonsuo appears to have rediscovered the Buddhist layer of 'concepts-constructed-in-relation-to-the-self-concept' which we discussed earlier.

Revonsuo makes an interesting suggestion concerning the *bizarreness* of dreams. In waking phenomenology, the possible *combinations* of phenomenological concepts are highly constrained by the configuration of the real world. We would not normally experience a friend turning into an animal.

¹¹ It was argued by some philosophers, such as Dennett (1991) that dreams are not conscious experiences, but artificial memories produced on waking. However, recent research has shown that they *are* conscious experiences. For example, Revonsuo (1995) cites experiments in which subjects are trained in lucid dreaming (the ability to realise that one is dreaming, and take control of the dream). When they became aware of the dream, they were able to give a prearranged signal to the experimenters, such as a small movement of their eyelids. This proves that dreams happen in real time.

But in dreams there is no such constraint; concepts are free to combine in such unusual ways.¹² But there may still be constraints on possible combinations imposed by the structural nature of phenomenology. Looking for common constraints in reports may thus provide clues about this structure. (Reports of hallucinations should provide similar data.)

We appear to experience the ‘real’ world using the same mechanisms as when dreaming (though with the addition of the more detailed low-level sensations). This leads to Revonsuo’s famous *virtual reality metaphor* of consciousness. Our phenomenological representation of the world is not necessarily structured in the same way as the real world – it is a representation constructed from concepts, dependent on our experience and goals, and informed only by our limited senses. For all we know, the real world could have a thousand dimensions, but we only have senses to detect three. When dreaming, we clearly construct a phenomenological world which bears little structural relation to reality. (The metaphor is put forward as an alternative to the ‘Cartesian Theatre’, in which a replica of the real world is reconstructed in the phenomenology.)

The VR lead him in turn to create what I call his *dream team* method (Revonsuo, 2001). This is proposed as a practical method for testing phenomenological/physical correlation theories, but I think that it also contains a gem of inspiration for a pure phenomenological method. The method involves four teams. The first is a set of *dreamers*. After their dreams, they each give a *verbal report* to an *animation* team, whose task is to construct VR simulations of the dreams as best they can. During dreaming, each dreamer’s brain activity is monitored by EEGs, fMRI or whatever technology is available, and this data is sent to a *scanning* team. The scanning team uses this data, together with their current theories, to attempt to construct a second set of VR simulations of the dreams. Finally, the two VR simulation sets are sent to a *judge* team, who attempt to match up the animation and scanning teams’ efforts. If this can be done successfully, then the scanning team is to be congratulated for having obtained a real, predictive theory of phenomenological/physical correlation.

Now, Revonsuo gives the credit to the neuroscientists, having taken for granted that the animators can reconstruct the phenomenology from the verbal report. He forgets that the animators are also faced with the apparently impossible problem of interpreting the reports. Recall that even if a formalised language, such as Marbach’s, is used, they can never be sure of the exact meanings that the dreamers assign to the *symbols* of the language. But Revonsuo has unintentionally come very close to providing a solution to this old phenomenological problem: if we allow the dreamers *themselves* to create their own VR simulations, then the descriptive problem is solved. There is no problem of deciphering the phenomenological meanings of *symbols*, because VR would reconstruct the phenomenology *itself* without the need for such *symbols*.

There would be refinements to make, however. We have seen that phenomenology can be principally constructed from *objects* rather than low level sensations – and Revonsuo’s VR method seems to imply that each pixel of sensation would be reconstructed. But this is not the case: in the phenomenology of the park dream, we see only a single, unitary *tree* object, and not its individual branches. We should not expect to produce a pixel-detailed representation of the tree – since that detail was not in the phenomenology. Instead, there should just be a single, unitary object, perhaps represented by a stereotype:



Revonsuo’s dream research has reinforced the Buddhist theory of a phenomenological layer of concepts, constructed in relation to a self-concept. He provides a metaphor of virtual reality as a way of thinking about how phenomenology is built from sense-data. And his dream-team method has given us a cue for a phenomenology method which escapes the problems of verbal language.

¹² Contemporary cognitive research on dreams (eg. Eichembaum, 2000) suggests that the function of dreams is to transfer a hippocampal cache of the day’s memories into systemised storage in the cortex. This involves associating the day’s concepts with similar concepts from memory. These associations may be played out in the phenomenology, and could perhaps explain the occurrence of bizarreness.

Object layers

It is interesting to note some further experimental evidence in agreement with the claim that we are usually conscious of high-level objects and not necessarily of low-level sensation. *Binocular rivalry* experiments (e.g. Lumer & Rees, 1999), present two different images are presented simultaneously to the subject's two eyes. The subject's phenomenology alternates between perception of one, then the other, image – even though the incoming sense data is the same. fMRI scans show correlations between 'higher' brain areas and the phenomenology; but not between low-level visual areas and the phenomenology. Crick and Koch (1995) conclude from such work that activity in V1 is not conscious.

Sachs' famous book (1985) tells of a neurological patient who mistook his wife for a hat. The patient could perceive the component *parts* of his wife clearly (so a Classical Introspection analysis would show nothing unusual in his phenomenology) but his object recognition system was damaged, giving an incorrect high-level object perception.

Experiments with visual inverting goggles (e.g. Welch, 1978, cited by Dennett, 1991) report that subjects adjust to upside-down vision within about a week – their high-level world becomes unaltered by the low-level inversion. However, if subjects pay close attention to their low-level perceptions (as in Classical Introspection and low-level Husserlian Phenomenology), they *then* report noticing the inversion.

Neuroscientist David Welchew (private conversation) has performed transcranial magnetic stimulation (TMS) experiments on himself, artificially stimulating parts of his visual cortex. He reports that stimulation of a particular high-level area produced triangles in his phenomenology, though on paying close attention he could tell that 'they were not really there'.

All of these results re-enforce the theory that there are multiple simultaneous layers of phenomenology, and that phenomenology can be focussed on any layer or layers without necessarily being focussed on the others. When we look at a painting of a tree, we may choose to see it *as* a painting; and note the canvas and brush strokes. Or we may choose to 'see through' that layer and conceive the tree itself. Or we may choose to pay close attention to our immediate lowest-level sensations, and see only the shapes and colors in our visual field. Or, more normally, we may see multiple levels at the same time.

From all this, we see that Classical Introspection was incomplete as a method for describing phenomenology. It artificially instructs its subjects to focus only on the lowest level, specifically ignoring the others. It may be a sound method for exploring this lowest layer, but that is all. It is a *special case* of phenomenology. James, on the other hand, made the opposite mistake of focussing only on the higher layers – when in fact there are usually both high *and* low layers present. The Buddhists are right about the layers – but perhaps it is unnecessary to 'peel away' the entire lower layer before being able to see the higher ones. Husserl seems to have been on the right track – beginning by describing the base, but also noting higher level concepts when they arise.

Some suggestions for future concept-layer research

It is interesting to ask 'how low can phenomenology go?': what is the lowest level of perception that is accessible to consciousness? The rivalry experiments show that activity in V1 is not *necessarily* conscious – one eye's image is always absent from phenomenology, even though it is present in V1 – but I disagree with Crick and Koch's claim that it is *never* conscious. Perhaps we *can* choose to bring its data in and out of phenomenology, by specifically focussing our attention upon our lowest level of perception in the manner of Classical Introspection? The rivalry result shows that we are not *normally* conscious of such a low level, but it does not rule out the possibility that we could become so by conscious effort. This would be a similar effort to that by which we can notice our visual world briefly go black when we blink. Similarly, is it well known that we normally perceive *differences* in color rather than actual colors: a white car still looks the same color under an orange streetlamp, because its relationship to its environment is unchanged. But could we, by deliberate effort, become conscious of the lower layer containing the original retinal data? Probing the depths of phenomenology would be an interesting research program, and I predict that trained Buddhist meditators may be able to dig deeper than normal subjects.

A common stage hypnosis trick is to give a pair of ordinary glasses to a hypnotised subject, telling him¹³ that they allow him to see through the clothes of the audience. The subject looks in amazement at the beautiful, apparently-naked women in the audience – but without receiving any low-level visual information from them. I suspect that the experience here is similar to that of dreams: the details will disappear if the subject focuses on them. It is only a high-level experience of the naked bodies; there is no low-level information! Bringing this trick into the laboratory (perhaps with a more politically correct scenario¹⁴) could be an interesting way to explore concept layers. What exactly happens when the subject tries to look at the details? Does he see blurriness (like in dreams), or clothes (like the inverted goggle subjects being able to see through the illusion when they concentrate), or do his higher-level concepts and memories try to ‘fill in’ these details (like false memories)?

Finally, rather than working from dream reports to construct and analyse the structure of dream phenomenology, it may be possible to use *existing* images for this purpose. Surrealist artists such as Salvador Dalí and Roland Penrose depict dream-like scenes: what do these pictures have in common? *why* are they ‘dream-like’? Again, the answer seems to be that the objects in these pictures are depicted as *stereotypes*, lacking low-level detail. Like a matchstick man, a stereotype is not a low-level depiction of a visual experience, but a *symbol* for a single, unitary phenomenal object. Perhaps by getting subjects to rate surrealist images for dreamlikeness, we could save the hassle of creating new images from reports, then work directly from these existing ones?

Recent fringe theory

Mangan (1993) reviews the contemporary view of James’ ‘fringe’ of consciousness. James’ original story was that it was a semi-conscious presentation of context information – specifically, of information about associations between the objects present in phenomenology and previously acquired knowledge. Mangan draws on recent psychological research into the capacity of ‘working memory’ (which, as Baars pointed out, is essentially a (limited) model of phenomenology), which is found to be very small; famously, ‘seven plus or minus two items.’ (Miller, 1956). Objects must thus be pulled rapidly in and out of phenomenology. But, as Crick and Koch noted, we don’t experience ‘tunnel vision’ of only seven sharply-focussed objects at a time. Instead, we notice a few objects in focus, and a general feeling of context around them. This ‘context’ resembles James’ fringe. When I turn my attention to these cups, I must shift the rest of the rest of the coffee shop out of working memory to make room for them. I am however still aware of the shop, but now in less detail. The coffee shop is in my fringe.

This process can be redescribed using the concept layers idea developed earlier. Initially, I am aware of the coffee shop at multiple levels of detail. I intend it as a whole *shop* object; I intend the *waiter* and my *table* as sub-objects; I intend the *head* and *body* of the waiter; I may even attend the low-level components of the waiter’s face. These approximately seven objects fill my phenomenology. But when I attend to the cups, I must remove many of the shop’s details to make room for them, and perhaps only intend the shop as a *shop*, without any sublayers. However, my SHOP concept contains *associations* which are still present in the phenomenology, and which allow me to recall the subcomponents when I need them. These associations are the fringe.

¹³ It is invariably a ‘him’ in this trick.

¹⁴ Though it may be easier to find volunteers for the naked women version.

Concepts

A new phenomenological model

“I should report that which I say I saw,
But know not how to do’t.”

- *Macbeth's messenger*

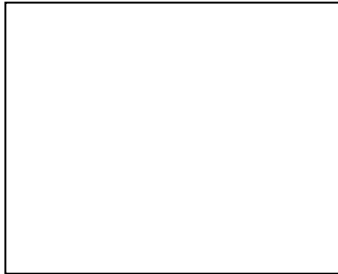
We have reviewed phenomenological research from a wide range of traditions: Introspection, Husserl, James, Buddhism, and contemporary cognitivism. We will here summarise our findings about the structure of phenomenology, and use them to develop a sketches of a practical model and a method for describing subjects experiences using the model. I like to think of this as a ‘neo-cognitive’ model: it is essentially a cognitive model, but one which is designed *explicitly* to model the content of consciousness¹.

The model will borrow ideas from theoretical Computer Science. The ‘serial digital computer’ (i.e. Turing Machine) metaphor is currently and justifiably outmoded in Cognitive Science, and many cognitive scientists think that theoretical Computer Science is therefore of little use to their cause. But they neglect that the latter is *not* merely concerned with information *processing*: it has also established concepts for thinking about large scale information *structures* - such as recursive data hierarchies and pointers - which have been mostly ignored by cognitive scientists. It is these *structural* ideas that we will draw upon for our model of phenomenological structure.

We have until now been unavoidably loose in our terminology, due to our surveying of work from radically different academic traditions. For example, we have used the words ‘concept’, ‘object’ and ‘noesis’ interchangeably. In the course of presenting the model I will introduce a single, unified terminology, and I make no excuse for mixing and matching from the different sources.

Pure consciousness

Pure, contentless experience is possible; it is possible to have a phenomenology without anything in it:



This is the simplest form of consciousness, and is reported to feel ‘joyous’ and ‘blissful’, suggesting that all other emotions are subtractive; joy is the *absence* of content².

¹ It is an old academic joke that ‘neo’ means ‘not really’.

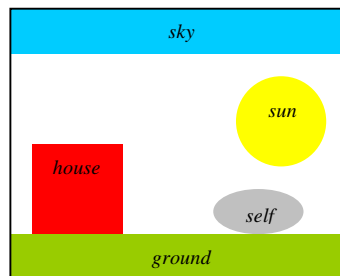
² As in ‘contents’ not ‘contentment’!

Moments

Consciousness is not continuous through time; it takes the form of discrete, instantaneous *moments* occurring at about 50Hz in humans.

The object world and the self

The usual form of consciousness is a phenomenology of *objects*. To *perceive* (or *intend* or *sense* – I use these words as synonyms) an object is to have it present in phenomenology. Human phenomenology is extremely limited in capacity, so only about seven objects can be present at any moment. An important object is the *self* which represents the body and history of the experiencer. However, there is nothing inherently unusual about the self – it is just one object of many. Human phenomenologies usually (though not necessarily) contain the self surrounded by a small cluster of other objects which currently concern the self. The set of concepts and their relationships is called the phenomenological *world*, and is spatial and three-dimensional. A child's phenomenology might contain objects representing his self³, his house, the sky and the sun; and their locational relationships relative to each other (note that children often perceive the sun to be beneath the sky):



(We already noted that in pure consciousness, phenomenology can exist without a world. Note now that the world can exist without a self object, or with a misplaced self-object, as is the case under the effects of hallucinogenics such as ketamine.)

Objects, concepts and nomena

Objects exist in the phenomenal world, not in the 'real', independent-of-human-observation world. This 'real' world only contains fundamental entities: quarks, quanta, a vat, the mind of God, or whatever they may be. It is beyond the limit of our epistemology to ever know for certain what is in the 'real' world (we can only hope for our theory to converge towards it). However, assume we have magically discovered that the 'real' world is ultimately made of things called *nomena*⁴. Now, there is no house in the 'real' world. There is just a collection of nomena arranged into a house shape, and nothing more. The house is something we invent to describe that set of nomena in a useful practical way. Houses, suns and selves *do* exist in our phenomenologies, but not in the 'real' world. (This is a standard scientific realist view, which we will formalise this idea at the end of this thesis (p.61). For now, I introduce it only informally.)

How, then, so we get a *house* into our phenomenology if our sense organs are only detecting the effects of nomena? Why do we not just perceive the low-level data from our senses? Where does the house come from? The answer is from our history: we have learnt to recognise the house pattern, and our

³ Neural and Psychological research into hippocampus place cells (e.g. Burgess & O'Keefe, 1996) suggests that there are two different modes of representing the self-object in the world. The *allocentric* mode is like that shown in the diagram here: the world is constructed, and the self is just one object in it. As the self moves around, the rest of the world stays in the same place. Secondly, the *egocentric* mode represents the self implicitly, as the perspective from which the world is constructed. In this mode, there is no self-object within the world, but as the self moves, the whole world moves in relation to it. The former is like a 3rd person video game: the world stationary with the character moving in it; the latter like a first-person driving simulator, for example. Hippocampal events corresponding to these modes have been found in rat hippocampi during maze tasks. Psychological evidence (e.g. Boroditsky, 2001) demonstrates the two modes in humans, based on their use of language: the phrases 'moving forwards' and 'moving backwards' can *both* be interpreted as 'moving towards me' or 'moving away from me' depending on whether I am in allocentric or egocentric mode. (Hence with a little thought, the Radiohead quote at the start of chapter 2 can be interpreted in at least three different ways – with the distance between the two protagonists getting smaller or larger or staying the same!)

⁴ I have invented the word 'nomena' to be reminiscent, though not the same as, Kant's 'noumena'. My nomenal world is going to be the counterpart of my *phenomenal* world. I should make it clear than I do *not* wish to import any Kantian baggage attached to 'noumenal', hence my fresh choice of word. My nomenal world is *only* what I have defined it to be. ('Nomena' should also not be confused with 'nominalism' – an entirely different philosophical concept.)

brain has formed a useful abstraction. This abstraction is called a *concept*. (I here follow the convention of the Cognitive Science concept literature of notating the concept of house as ‘HOUSE’ and the house object as ‘house’.) Once we (unconsciously) possess the concept HOUSE, we can then create *instances* of the concept in working memory – these are *house* objects. It is these objects that reach the phenomenology – we are never conscious of concepts themselves. (This formulation is borrowed from the Object/Class distinction formalised by Computer Science.) Concepts are essentially Plato’s forms – they do not exist in our phenomenal worlds, but they are the templates for the objects that do.

A model of concept structures

We have seen that the phenomenal world is comprised of objects, which are instances of concepts. So to study the structure of the phenomenal, we need a theory of the structure of concepts. I here present my own formalisation of concept structures, based on philosophical and psychological research. (A preliminary argument building up to this model can be found in (Fox, 2001)). This is not a complete theory, rather a sketch which I hope is sufficient for our purposes here. Specifically, we are not here concerned with the *mechanisms* of concept formation and computation; only with their *structure*. (Though as we will discuss later (p.85), fleshing this theory into a full computer model could perhaps eventually be part of the path to building conscious machines.)

Overview

The theory is again borrowed from Computer Science, this time taking the idea of *primitive recursion*. A primitive recursive system is one in which all possible entities are constructible from a small set of primitive atoms and a set of operators for combining them. The atoms of the concept system are to be low-level, neurally-hardwired concepts. Concepts can be combined with three operators (each given a symbol) to form new concepts: *generalisation* (λ), *composition* (+), *analogy* (\leftrightarrow). I will introduce the theory first with its historical background, then in informal language, then with the formal notation.

A brief history of concept theories: definition vs. exemplars

Descartes (1641) believed that some concepts are innate, hardwired into our brains, such as GOD and INFINITY. This was because there is apparently no other way of obtaining such concepts. Locke (1690) rejected this, comparing the newborn mind to ‘white paper, void of all characters, without any ideas’. He refuted Descartes by showing that there are people who do not possess the supposedly innate concepts – such as children. Children do however have the capacity to *learn* those concepts. Locke makes a distinction between ‘sensible qualities’ such as colors and sounds, and concepts (which he calls ‘ideas’). Lockean concepts are formed by exposure to sensible qualities, although he does not specify a mechanism or structure for this process.

Hume (1748) builds on the Lockean view and puts forward a theory of concept *structure*. Concepts, he says, are complex. Basic concepts are formed by sensation (of sensible qualities), but concepts can be combined to produce more complex concepts. We get the concept MAN by seeing men, and the concept WINGS by seeing wings. But we can form ANGEL by combining the two – without the need to see angels. Note that there are two processes involved in Hume’s picture: concept extraction through *generalisation* (e.g. learning MAN by generalising from seeing four men), and *composition* of two or more concepts to form a new one. He does not discuss mechanisms for these processes.

Echoing Descartes, Kant (1783) tried to resolve the tension between the innate-concept school and the ‘blank-page’ school, by suggesting that *some* innate concepts are required in order that the Humean *mechanisms* can take place. These concepts were to include CATEGORY, CAUSE and SUBSTANCE.

The notion of complex concept formation through composition and generalisation was attacked by Wittgenstein (1953), who denies that definitions exist for any concept. Rather, the meaning of words is defined by their *use*; that is, by the set of all past experiences of them. This is a little like the Humean generalisation idea, but suggests that generalisation occurs at the moment of word use rather than during concept formation. For example, if Hume was taught the concept MAN by exposure to ten pictures of men, he would then generalise their common features, so MAN would perhaps be defined as a MALE PERSON, and this definition would be stored in memory. Wittgenstein, on the other hand, would store *every instance* of men that he had seen, and refer to *all* of them when deciding when to use the word ‘man’ in future.

The history of concept structures has thus been one of two tensions: between innateness and learning; and between structural definition and exemplar usage. These tensions have continued into contemporary cognitive models.

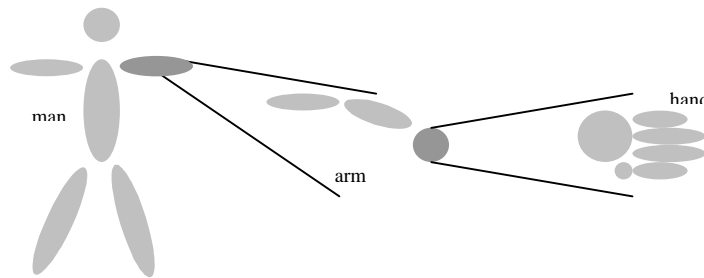
Contemporary structural models

Humean concept composition is embedded in our everyday thinking – we think of concepts as composed of subconcepts. Computer programming languages have been steadily evolving over the past 50 years, bringing their structures closer to those of human thought to aid their use. The current generation, *object-oriented* languages, formalises composition perfectly – Hume’s angel could be written in Java as:

```
class Angel {
    Man m;
    Wings w;
}
```

The best-known decomposition theory of *action* concepts is Schank’s (1972) Conceptual Dependence Theory, in which the semantics of verbs are broken down hierarchically into (roughly) 12 primitive actions. These actions include PTRANS, the transfer of physical objects, MOVE, the movement of a body part, and INGEST, the intake of food or air. (These primitives could be innate, or learned in infancy). For example, RUN might be a combination of MOVE, PROPEL (applying force to a physical object) and PTRANS. According to CDT, even a verb such as ‘to program’ is ultimately defined in terms such as body movements, ingestion and excretion!

In image and object-recognition, Marr (1982) suggests a hierarchical representation of *physical* concept and object structures. E.g. the human body as a series of cylinders at different levels of generality:



Thus, we can recognise an instance of MAN by recognising cylinders arranged in appropriate locational relations. To a first approximation, we don’t have to look at the internal structure of the ARM (for example) – working on the MAN hypothesis, we just assume that the cylinder in the arm position is an ARM, and check whether it is in the correct position in relation to the others. Later, we may inspect the putative ARM more carefully, to check that it really is an arm. If we are anal enough, we may further check that the HAND is really a HAND and so forth – but eventually we just assume that some small-scale component really is what we think it is, given the substantial match between the arrangement of the high-level structures above it and our prior conception of MAN.⁵

Contemporary *Categorisation* research (reviewed in Eysenck & Keane, 2000) models how we decide when one concept is similar to, or an instance of another: is WAR a GAME? are HOUSES and SHOPS similar? Structural models of categorisation use *definitions* of concepts in terms of other concepts to make these decisions. These come in different flavours: *classical definitions* define each concept as the conjugation of subconcepts (e.g. An ANGEL is a MAN plus WINGS); *prototypes* are less strict, providing a set of subconcepts which are typical but not necessary for the defined concept. (e.g. A

⁵ The alert reader may notice an apparent oddity here: that no matter how hard we look, we can never actually be sure whether we are looking at a MAN or not. For it may turn out that at some level of detail, some subcomponent such as the blood cells has been replaced by something else, such as silicon blood cells. This would have a knock-on effect: the hand would no longer be a HAND, since it is not made out of blood cells, as the definition of HAND requires; so the man would not be a MAN. In practice, we do not bother to check the matches of potentially near-infinite levels of details of definitions: when we see correct matches up to some level of detail, we *stop* and simply *assume* that we have correctly identified a MAN. This argument is an interesting prelude to my argument for ‘how to identify consciousness’ that is presented later in this thesis (p.76).

CAR usually has four WHEELS and an ENGINE and a DRIVER – but a three-wheeled car would still pass this definition as it is similar to it, though not being an exact match.)

Contemporary exemplar models

Wittgenstein (1953) suggested that definitions are not stored as extracted sets of subconcepts, but simply as sets of exemplars of use of the concept. Recent cognitive models (reviewed by Ramscar & Hahn, 1998) have implemented this idea, and make category judgements of a new instance based on its average similarity to *all* the previous instances. The justification for this view is that given any definition of a concept, it always seems possible to find a counterexample which defies it. (Wittgenstein famously demonstrated this with the concept of GAME.)

A synthesis

There are elements of truth in both exemplar and composite definition theories – and I think the two theories can be easily combined as follows: we initially learn concepts by experience of exemplars. Each exemplar is perceived as a collection of low-level objects. Initially, the set of these exemplars is stored as the concept definition. My definition of HOUSE would be the set of all houses I have seen.

However, over time, the exemplars become gradually ‘digested’ into composite feature definitions. My memories of the individual houses begins to fade, to be replaced with a more generic, stereotyped concept of HOUSE. This make take place over months or years. We are all familiar with this feeling: my memories of walking down the Royal Mile in Edinburgh this year are fairly distinct; but my memories of walking down Kings Parade in Cambridge as an undergraduate have sadly faded, replaced by a single generic memory of that place, devoid of each instance’s individual details.

This theory is backed up by psychological and neurocomputational modelling research on memory (e.g. Eichembaum, 2000; McClelland *et al.*, 1994; O’Reilly & Rudy, 2000), which suggests that recent memories are stored relatively unprocessed in the hippocampus. During the night, they are gradually processed and stored in more abstract ways in the cortex. This could possibly be an explanation for the bizarreness of dreams: features are extracted from the recent memories and joined with existing cortical concepts. This joining of old concepts with newly acquired information could explain why dreams often feature the day’s events joined in bizarre ways with stereotypical objects.

Piagetian learning theory also fits this theory (e.g. Keil & Batterman, 1984). Children first learn concepts by exemplar, then famously form over-generalised rules. Finally, the rule exception cases become accounted for. Again we see feature patterns extracted from exemplars over time. (It is an interesting whether the exceptional cases are added to definition as exemplars or by new feature definitions: I suggest that initially they are exemplars which are again gradually generalised as far as possible.)

We have seen that concept definitions appear to be stored as combinations of *both* exemplars and subconcept features, and are gradually compiled from the former to the latter. I claim there is also a third and final component of concept definitions: *analogy*.

There are some concepts which we cannot form from any exemplar and composition combinations of our low-level input concepts. For example, the concept of ATOM and other non-sensible concepts. We are typically taught what an ATOM is by analogy to the solar system: the electrons are to planets as the nucleus is to the Sun. In fact most abstract scientific concepts seem to be formed this way – compare how we introduced Dennett’s Multiple Drafts Model and Baars’ Theatre of Awareness earlier in this thesis by drawing analogies with macroscopic concepts.

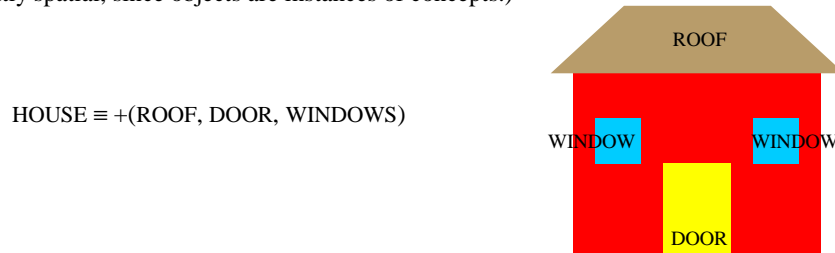
Formal notation

The HOUSE concept mentioned above is initially learnt and stored as a set of *exemplars* of houses⁶, where each exemplar is perceived as a collection of instances of previously acquired concepts:

$$\text{HOUSE} \equiv \lambda(\text{experience_of_cottage}, \text{experience_of_villa}, \text{experience_of_flat})$$

⁶ The ‘ λ ’ symbol is borrowed from λ -calculus, where it *very* roughly means ‘abstraction over’

HOUSE may become transformed into a feature-composite concept, comprising ROOF, DOOR and WINDOWS in a particular configuration. These feature are gradually extracted over time from the λ -definition. We should really draw a picture to show the spatial relations of the subconcepts, but I also an the abbreviated notation for illustration purposes: (Concept space, like phenomenal object space, is inherently spatial, since objects are instances of concepts.)



(On relatively rare occasions we are given an explicit definition of a concept – such as in mathematical definitions. In these cases, the exemplar stage is not needed and the composition representation is formed immediately⁷.)

Finally, a few concepts are learned and stored by *analogy*. Most children are taught the ELECTRON concept by analogy to the solar system⁸:

$$\text{ELECTRON} \equiv ((\text{SUN, PLANET}) \leftrightarrow (\text{NUCLEUS, ?}))$$

(The question mark indicates the place of the new concept within the analogy. This notation is borrowed from unification theory in computer science.)

Using existing concepts and the three operators, more and more complex concepts can be constructed. After acquiring HOUSE and other component concepts, we could go on to form VILLAGE, CITY, COUNCIL, ELECTION and so on.

An example of concept formation

Suppose we are learning the concept CAR by being shown cars, and that we already possess the concepts DRIVER, WHEEL, RED, BLUE, FAST, SLOW and WINDOW. We are shown three cars, and we perceive each of them as a collection of instances of the latter concepts. After perceiving each collection, we can store a concept representing that collection:

$$\text{CAR1} \equiv +(\text{DRIVER, WHEEL, WHEEL, WHEEL, WHEEL, FAST, BLUE, WINDOW})$$

$$\text{CAR2} \equiv +(\text{DRIVER, WHEEL, WHEEL, WHEEL, WHEEL, FAST, RED, WINDOW})$$

$$\text{CAR3} \equiv +(\text{DRIVER, WHEEL, WHEEL, WHEEL, SLOW, GREEN, WINDOW})$$

(recall that the ‘+’ is shorthand for spatial locations.) Our new CAR concept is thus initially this set of exemplars:

$$\text{CAR} \equiv \lambda(\text{CAR1, CAR2, CAR3})$$

Over time, the mechanism extracts the features. This is done gradually, so at some intermediate point the definition might be: CAR \equiv + (DRIVER, WINDOW, $\lambda(\text{CAR1, CAR2, CAR3})$)

⁷ However, even mathematical concepts learned originally by composition tend to later absorb exemplars into their definition. Suppose the concept PI has been taught as some composition of CIRCLE, DIAMETER, RATIO. The student then performs many calculations with this concept, and ‘gets a feel’ for how it is used and what it is for. If we then introduce the student to non-Euclidian geometry, the circle ratio now changes! Should PI change its numerical value to the new ratio, or stay the same? Depending on their previous experience with PI, students will form different answers to that question. There really is no definite answer – it is matter of choice how to carry the concept over to the new domain. Most maths books retain the numerical concept, but some (such as *The Adventures of Mr Tompkins*) allow the value to change.

⁸ A more controversial example is how one can conceive of the VAT in Putnam’s ‘Brain in a vat’ scenario. *Prima facie*, it may appear that if one is a brain in a vat, one can never conceive of the vat because one obtains no sensations of it from which to begin constructing concepts. But by defining the concept using analogy the problem is solved. One simply has to conceive of a brain in a vat in *this* world, then draw an analogy between the virtual world of the brain and ones own world. This allows the ‘meta-vat’ to be conceived as follows: META-VAT \equiv ((VIRTUAL_WORLD, BRAIN_IN_VAT, VAT) \leftrightarrow (MY_WORLD, MY_BRAIN, ?)).

And eventually: $CAR \equiv + (DRIVER, WHEEL, WHEEL, WHEEL, WINDOW)$

(Perhaps each term should also have a weight, to allow the λ -term to gradually fade out.)

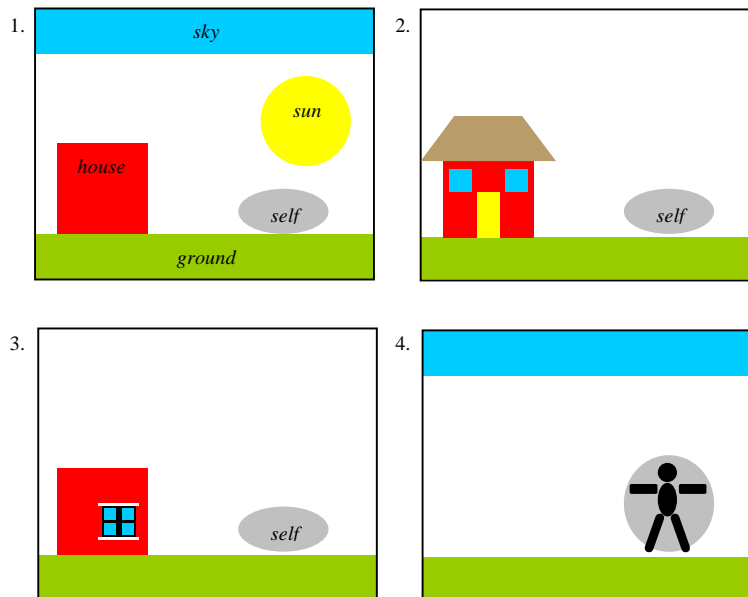
A model of phenomenal objects

Now, what does all this have to do with our phenomenology? Quite a lot – because the structure of phenomenology is essentially a structure of objects, and objects are just instantiations of concepts, so share their structure.

Multiple intentionalities and the fringe

We have seen that objects have a hierarchical structure: a *house* can be decomposed into subobjects (*door*, *window*, etc.); and each of those can be decomposed into subobjects, and so on, until we eventually reach memories of sensations. Now, recall that phenomenology has a very limited object capacity – only about seven objects can be present at a time. If we instantiated the entire hierarchical structure of *HOUSE* when we saw a house (such as in the house, self and sun diagram earlier), then there would be little space left for any other objects! Computer Scientists are very familiar with this problem, and have devised a solution: *pointers*. Rather than instantiating the whole structure, a single *house* object is created. The object does not contain the concept's structure, but contains *associations* to the concepts from the single next layer of detail. So a *house* would contain associations to *DOOR*, *WINDOW* and *ROOF*, but not to *their* subcomponents or subsubcomponents. Just three associations are brought in. If more detailed information about the house is needed, only then will the subcomponent objects be brought into the phenomenology. Then, if more detail about the *window* is needed, its subcomponents can be instantiated and so on. Each of these instantiations is of course at the expense of an old object which must be removed from the phenomenology to make room for it.

The following diagrams show the effects on phenomenology of focussing on different parts of the same scene. (1) attention spread evenly; (2) focussing on the house, (3) 'zooming in' on a window, and (4) focussing on the self. Note how the detail of other objects is reduced as the object capacity is used for the focussing (The number of objects is limited to seven plus or minus two):



My use of the stereotypical children's drawing scene here is deliberate. Children tend not to draw scenes as presented by their low-level senses: rather, they depict their actual phenomenology, using *symbols* to represent each unexpanded object. The sky is drawn as a blue bar at the *top* of the page, rather than extending down to the horizon as 'realism' would recommend, because the child perceives the sky to be at the top of the world. Matchstick men are used as symbols to represent instances of the non-expanded PERSON concept; and the stereotype 'house' symbol (door, two windows, chimney with smoke coming out) represents almost any kind of house. The child later loses this remarkable ability to

accurately depict the contents of phenomenology - presumably when indoctrinated with western realism in school art lessons. Great artists learn to go back to depicting high-level unexpanded concepts using similar symbols; Magritte's pipes being a classic example. This allows more accurate communication of concepts without danger of confusion or distraction by irrelevant details. The *conceptual art* movement can be seen a deliberate attempt to restore this lost form of pure concept communication. (So it is not surprising to see the influence of Zen Buddhism on great conceptualists such as Cage (e.g. 1939)).

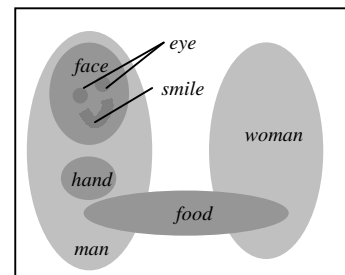
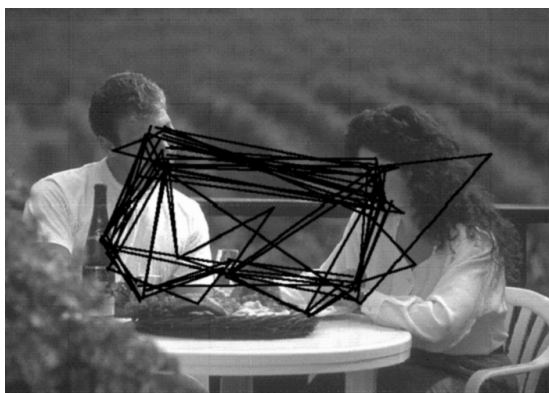
This process of expanding associations captures James's notion of the fringe: the vague semiconscious information which becomes vivid and conscious when one tries to look at it. Just as in Computer Science: pointers are not actually *data*, but a means of getting to data; we might call them 'semi-data'. The process also matches Crick and Koch's 'spotlight of attention' theory, and shows how and why Husserl's multiple layers of intentionality come in to being.

Depending on the preconscious effects of hormonal activity (for example), the spotlight may be influenced to bring certain aspects of objects into phenomenology rather than others. We may focus on the uglier aspects of a woman rather than the beautiful ones. We might also perceive her as an instance of a Jungian ANIMA concept, rather than simply as a WOMAN. These are examples of James' *moods*: we perceive the world in different ways, producing different feelings.⁹

Objections to the model

It may be objected that intuitively we feel there are a great many more than seven objects in our phenomenology - surely we can see a rich, detailed, colourful world in front of us? But as Dennett (1991) points out, this 'Cartesian Theatre' of a fully re-constructed world in the mind is an illusion. A better explanation is that, like the eyetracker experiment mentioned earlier, it *appears* to us that the detail is there. If we examine any *part* of the scene and ask ourselves how much detail we can see, the answer is 'lots' - but that is precisely because we have focussed on that small part in the process. Those who think they have such 'photographic' phenomenology are advised to play the old party game: ten random objects are placed on a tray and shown to the players for a few seconds. They then have to remember them - to make a phenomenological report, in our technical terminology - and rarely manage this feat. Detailed perception of the kettle (say) on the tray is only possible at the expense of losing focus on the others. If, as Husserl showed, memory is the replaying of experience, then photographic phenomenology is as unlikely as photographic memory.¹⁰

Recent results (e.g. O'Regan, 2001) from *change blindness* experiments hammer the point home beyond dispute. Two versions of a photograph are alternately flashed to the subject in quick succession. The versions differ in that one features a movement or removal of a large but *unimportant* object. (The second version is obtained by airbrushing the original photograph). For example, the fence in the picture below left (ignore the scribbles for now) was moved about 1cm higher in a second version (see O'Regan, 2001, for the two flashing versions):



⁹ Very little work has been done on emotional phenomenology - see footnote on p.39.

¹⁰ Investigation of the phenomenology of those rare individuals who have the 'photographic memory' capability would be interesting. I predict that their memory is not really of the precise low-level sensations implied by the name; rather that they just have an especially large phenomenological object capacity - perhaps of around a hundred objects rather than seven.

Subjects are asked to find the difference. But even when the images are exchanged at rates of several times per second, they are unable to find it for up to several minutes of intense searching. This is because they do not usually include the fence in their phenomenology: only objects corresponding to *important* parts of the image are created. For normal subjects, such objects include the facial features of the couple, the hands, and the food. Furthermore, subjects' eyes rarely even *look* at the fence: the 'scribble' in the above picture is in fact an eyetracking of a subject searching for the change; his eyes just keep scanning the parts of the picture corresponding to his phenomenology, which probably has a similar form to diagram on the above right.

A second objection to the model is that the story we have told so far seems only to involve *visual* information. Where are the sounds, smells and relations? The sense questions are easy to answer, the relations question is harder. With regard to senses, the model does *not* claim that the phenomenal world is *visual* – we have just been using visual diagrams to discuss it. It is *spatial*: it locates object in three dimensions, but it is not a 'mental image'. It is *objects* that are located in the space, not *images*. Our concepts contain a mixture of sensory information: a subcomponent of our TROMBONE definition may be a set of exemplars of trombone sounds we have experienced (or a *composition* of generalised sounds, of course). Our discussion has focussed on visual aspects purely because they make the diagrams easier. (For example, Klawiter (2001) demonstrates that natural audition functions similarly to vision: its purpose being to provide information about objects in space¹¹ – and *not* usually to present the low-level sounds themselves to consciousness. Listening to music, like looking at paintings, is a special case where we do choose to focus on the lower-level objects.)

I know of two possible answers to the relations question. (Examples of the question include 'Where is love in the phenomenal world?' and 'How is the fact that John loves Mary expressed in that world?'). The first answer, which I find unsatisfactory, is from James, who postulates that relations are non-phenomenal. All processing of relations is done preconsciously, and consciousness is only presented with the *results* of that processing. James appears to reach this conclusion by default: he can think of no way of fitting them into the phenomenology, so therefore they must be unconscious. I find this unsatisfactory for two reasons: First, it multiplies our ontology: there are now two different kinds of objects; conscious and nonconscious ones. Second, there is a distinct phenomenal feeling of what is it like to know that John loves Mary. We consciously perceive John and Mary in a different way if we know of their relationship. *Something* must change in the phenomenology.

The second, better, solution is just to allow relations to be objects. Remember that *objects are not images*. So our definition of JOHN could comprise concepts of TALL, BLONDE and LOVE_FOR_MARY.¹² This concept can be projected into phenomenological space just like any other. Introspectionists have historically had difficulty with the existence of such concepts precisely because they made the false assumption that everything in phenomenology was *sensation* rather than *conceptual* objects.

I would like to consider how our model responds to the infamous 'imageless thought' debate. When, for example, we image a rotating triangle, is there actually a triangle in our phenomenology? Except in rare hallucinatory patients, there is obviously no triangle as vivid as if we were to hold a paper triangle in front of our eyes. But I claim that there *is* a *high-level* triangle object there – it does feel like something to imagine the triangle, and I *do* locate it at a particular place in the world as I imagine it. The situation is like dreams, but more so. Revonsuo notes how dreams are less vivid than reality, they cannot be focussed on at low-levels because there is no low-level sensation present, only high-level objects. In the same way, when performing an abstract geometric task (rotating the triangle), I have no *need* to perceive it at any level lower than necessary for the task. I perhaps represent the lines and corners, but no deeper. A dream triangle may be yellow, but not show any surface texture; an imagined triangle does not even have color. Interestingly, Buddhist legends hold that one who has attained a very high level of mental control can actually *make changes to the world*, such as creating objects.

¹¹ Klawiter's experiments typically involve listening to stereo-pair recordings of, say, a car driving past the subject from left to right, on headphones. The subject becomes aware of the location of the car in space, and not necessarily of the detailed sounds it is making. Klawiter (private conversation) is critical of traditional psychological sound experiments which play monophonic headphone sounds to subjects, as this is a very unnatural situation from which it is difficult to make natural generalisations. (In these special cases the sound may not appear to be localised, except perhaps as 'inside' the listener's head.)

¹² How do I acquire the LOVE_FOR_MARY concept if I have never had the experience of loving Mary? By analogy: LOVE_FOR_MARY \equiv (LOVE_FOR_JENNY, JENNY) \leftrightarrow (?, MARY), where Jenny is the one whom I love. (Those who have never loved *anyone* will not be able to make this analogy, so will be unable to form LOVE_FOR_MARY.)

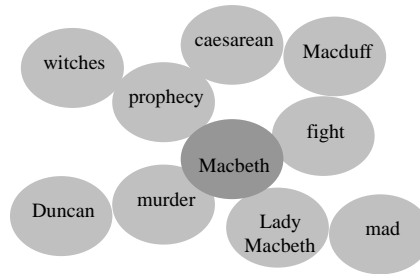
This could be evidence that although the normal mind has no need to create detailed low-level hallucinations - so therefore does not normally possess the skill – it could still perhaps be learned by such mentally agile individuals.¹³

Phenomenology of abstract objects, exemplars and analogies

We have already shown that it is possible (in fact common) to perceive high-level objects without also perceiving their low-level subcomponents. We have also noted that ‘relations’ are just a kind of high-level concept, whose instances are perceived in phenomenal space like any other object. When we perceive John loving Mary, we perceive an instance of the LOVING_MARY concept in space.

Everything that we perceive is in phenomenal space – even instances of supposedly ‘abstract’ concepts. There is of course no borderline between ‘abstract’ and ‘non-abstract’ concepts; rather a scale from low- to high-level concepts: BLACK, STRAIGHT_LINE, TRIANGLE, DIAGRAM, GRAPH, EQUATION, EQUATION_WHICH_LOOKS_INTERGRATABLE_BY_PARTS...

Even when we are engaged in our most ‘abstract’ thinking, such as solving integrations, we still just perceive instances of concepts in space. Somewhere out there in front of us is an INTEGRATION. Of course, we do not ‘see’ anything *visual*; like the rotating triangle, this perception is of a high level object without lower level subcomponents. Similarly, when we recall the plot of a Shakespeare play we may summon phenomenal objects in the space before us. This is why spatial diagrams help greatly in comprehending such abstract structures – they aid our phenomenalisation of those structures¹⁴:



It is well known that the spatial and visual areas of the brain are used during ‘mental imagery’ tasks such as imagining motion on a map (Kosslyn *et al.* 1978). I hypothesise that the spatial areas are also used for *all* kinds of high-level reasoning.

The objects used in all of our pictorial examples so far have been instances of *compositionally-defined* concepts. A question related to that of abstract object perception arises when we ask how we perceive instances of *exemplar-defined* and *analogy-defined* concepts. What enters my phenomenology when I imagine and try to focus on an electron in front of me, if I have formed my ELECTRON concept by analogy to the solar system? The answer, of course, is that the elements of the analogy are present on the fringe of the concepts, and are brought into my phenomenology when focussed on – just as for the elements of composite concepts’ definitions. Hence when I think in detail about a physics problem involving electrons, I do actually perceive the problem in terms of ‘little planets’ in front of me – though yet again these perceptions are at a very high level, and are not ‘visual’ perceptions.

What happens when an exemplar-defined object is focussed on? What if my HOUSE is a set of exemplars of houses, and I try to perceive a *house* object? Yet again, I will bring the elements of the definition from the fringe into phenomenology – in this case, one or more of the exemplars (depending on how hard I focus); and yet again these will be high-level, not ‘visual’ perceptions.

Removing the Humean sense/object distinction

We have so far been a little cagey about where the lowest-level concepts come from in the model. At first reading of theory, it may appear that there is an ontological distinction between ‘atoms of

¹³ Salvador Dalí also claimed that he could summon hallucinations at will.

¹⁴ The author often draws such diagrams when needing to rapidly get up to speed with a new Shakespeare play – they are usually drawn on the programme whilst reading the programme notes, minutes before the curtain rises. They’re also useful for Wagner operas and anything else with potentially incomprehensible dialogue.

sensation' and 'high level objects' - that objects are not sensations and that sensations are not objects. This distinction goes back to Hume's (1748) 'impressions' and 'ideas'. *I wish to completely remove this distinction.* I treat the words 'perception', 'sensation' and 'intention' as synonyms, and all perception in my model is of objects only.

Experiences of objects *are* sensations, in *just the same way* as experiences of redness are sensations. Redness is a concept, just like other concepts, and we perceive instance of it in just the same way as any other concept. This claim may need a little justification:

It is time to banish the folk formulation of the phenomenal as comprising the 'five human senses': sight, smell, hearing, taste and touch. Taught to all of us as children, it imposes rigid distinctions between five discrete sense modalities. In reality, information from our senses can blend into one another before reaching the phenomenal: e.g., we cannot untangle our perceptions of taste and smell when drinking wine. The existence of synaesthiacs shows that they can become truly blended. Further, reports from recently-blind users of pseudo-vision devices (such as tactile displays mounted on the back's flesh), show that the sense-distinction cannot be made – users typically dismiss the question of 'which sense does it feel like, vision or touch?' as meaningless. (e.g. Dennett, 1991). The brain receives input information from many sources, and some of this information ultimately causes objects to appear in the phenomenology. But there is no real distinction to be made about what 'sense modality' the objects have. If further evidence is needed, recent work in brain-computer interfacing (reviewed in Vaigan, 2000) has fed information *directly* to and from neurons (using invasive electrodes), completely bypassing the senses. Again, users are consciously aware of the sensations, but refuse to modally classify them. Using my cocktail party head-up display mentioned earlier would not necessarily produce a *visual* experience – rather it might just make the user conscious of the information itself; if my guest is a trombonist, I might just perceive a high-level *trombone* object. Whether neural activity is caused by eyes, ears or electrodes is irrelevant. An experience of redness is just a phenomenal object, regardless of whether it was caused by the eye or by artificial brain stimulation; and an experience of a car is just a phenomenal object, regardless of whether it was caused by the eye or the ear.

We now see why sensation of a triangle, say, is not inherently different from that of redness. Somewhere in the visual cortex is, presumably, a neuron which fires in the presence of triangles. And if we are paying attention to it, its triangles appear in the phenomenology. In the same way, if we pay attention to the detection of redness (assuming we can reach that low level), then that sensation appears in the phenomenology. Redness and triangles are both just sensations. We presumably have neurons (or assemblies of neurons, or γ -oscillation-bound collections of assemblies) which learn to activate in the presence of high-level concepts such as GRANDMOTHER; and their phenomenal objects are much the same as those produced by lower-level detectors for concepts such as light or color¹⁵.

We can also note the blurring of the Humean distinction by considering of the lowest 'depth' that people can perceive. We have seen that some people can dig deeper than others: such as trained meditators and introspection subjects. *There is no universal lowest level of phenomenological access that could be called 'atomic'.*

Moments and phenomenology structures

We now have a description of the structure of a static phenomenological state, as containing a *prima facie* surprisingly small number of objects. Each object is an instance of a concept, and is a *perception*. (Or 'sensation' or 'intention' – these are synonyms). Concepts may be as simple as the detection of redness, or the detection of a complex hierarchical feature structure and/or similarity to exemplars. Objects in phenomenology contain associations to their subcomponents, allowing more detail to be brought into focus if required. The illusion of high-resolution everywhere is given, because as in Dennett's eyetracker experiment, as soon as one thinks about a part of the world, it comes into focus.

We noted earlier that phenomenology comes in instantaneous moments, occurring at a rate of about 50 per second in humans. As they are *instantaneous*, there can be no change in the phenomenology

¹⁵ There has recently been much philosophical discussion over whether low-level sensations are concepts (e.g. Kelly, in press). Such discussions seem to assume the discrete distinction between the five senses. The whole debate vanishes into a matter of terminology under our new formulation of the continuum of sensation. I class all sensations as concepts.

during a moment (because there is no ‘during’). Processing must therefore be done non-phenomenological, between moments. What is presented in the phenomenological moment is the current *result* of this processing. Dynamically, then, our phenomenological experience consists of a series of still frames of phenomenology presented at about 50Hz. (Conveniently, this is approximately the same rate as frames of cinema. This should be no coincidence: the cinema rate is chosen *because* it is the minimum required to feed us with a new frame for each of our moments).

A new method for phenomenology

Unlike most phenomenologists (e.g. those writing in the *Journal of Consciousness Studies* special phenomenology issues: Vol. 6. Nos. 2-3; Vol. 8. Nos. 5-7.) who merely *discuss* phenomenology, we have proposed a *model* of the structure of phenomenology. It has been intentionally designed¹⁶ to be easy to simulate computationally. Note, this is *not* to imply that a simulation would be conscious; only that it would contain the same information structure. It is easy to imagine a computer framework that would allow us to rapidly construct models of individual phenomenologies, if we had a method for extracting the information from the subjects.

We found earlier that there is a problem in performing such extraction with traditional verbal reports: the subject and scientist cannot agree on the exact meanings of the words used – even if a specially formalised language is used. But building on Revonsuo’s insight of updating extraction with modern technology, we could allow the subject *himself* to use our computer system to model his *own* structure. The scientist would not need to understand the subject’s concepts to figure out the object structures in the report – since the subject reports his own structures without assistance.

I am not an experimentalist, but here is a rough suggestion for how such a system might be used in an experimental protocol: For a brief time period, we flash at the subject a scene containing objects with several levels of detail. For example, there might be a house, made up of doors and windows, made up of smaller components. We provide the subject with a graphical interface to our reporting system, which contains stereotypical symbols corresponding to all of the kinds of object in the scene. The subject then drags them around to reconstruct as much of his phenomenology as possible from his moment of exposure to the scene.

None of this is to say that the scientist would then know what the structures *feel* like. The subject may draw an object labelled ‘cravat’ and an ungentlemanly scientist may have no idea what this *means*, not possessing his own CRAVAT concept. We have not solved the Nagellian ‘what is it like?’ problem, but we do not need to. The point is that we have obtained a full description of the *information structure* embedded in the phenomenology. And this is precisely what we need for our project of finding correlating information structures between the phenomenal and objective.

Once we have the raw information from the report system and from the objective recording devices (EEGs etc.) we can then apply statistical techniques to extract the best recorded approximation to the transitive correlate.

¹⁶ No pun intended.

Explanation

Bridging the gap

“It is nonsense, says reason.
It is what it is, says love.”

- *Erich Fried*¹

Our discussion so far has been a march towards correlation – toward a meshing of data from the first and third persons. In the third person we descended from coarse-grained behavioral and functional approximations of the creature correlate, down through detailed neural and subneural approximations, and eventually to the quantum level; hoping to home in on the transitive correlate and to see the structure of the phenomenal revealed before us in the objective. We constructed a model of phenomenology, which (if refined and employed empirically) will provide us with this same structure but obtained in the first person. Once we have empirically found the transitive correlate, we should then be able to find the true creature correlate acting upon it, transferring the structure from the physical to the phenomenal.

This chapter addresses the question of what we should do once science has found the creature correlate. How can we go beyond the *correlation* to form a *theory* of consciousness? (This process is represented on the cover picture by the dotted lines bridging the two structures.) Could we *identify* phenomenal objects with, say, the activity of GABA receptors? Or with 40Hz oscillations, or quantum state reduction? Or will this identity prove to be beyond the reach of our science or our concepts?

From Correlation to Prediction

A key distinction is that between *identity*, *correlation* and *causation*. Velmans (2000) provides a neat table of the differences. Here, *symmetry* means that if A correlates with B, then B correlates with A; and obeying *Leibniz's Law* means that if A correlates with B, then A and B have the same properties:

	<i>Symmetry</i>	<i>Leibniz's Law</i>
Causation		
Correlation	✓	
Identity	✓	✓

For example, imagine a house with several rooms, some of which are heated. Heat can be *caused* by the action of turning on the power switch of a radiator, but is not identical to, or correlated with, that action (since it may also be caused in additional ways, such as lighting a coal fire). Suppose that we repeatedly observe the house's inhabitants, and always find them in the heated rooms. In this case, the presence of people is a *correlate*, but not a cause or identity, of the heat. Of course we know that heat is really *identical* to the mean kinetic energy of molecules. Now, which of these three relations is necessary for forming a theory of heat; or for a theory of consciousness?

Before answering this question, we should remind ourselves of *why* we are trying to understand consciousness. Recall that this thesis is motivated by a list of *practical* questions. Our objective is to

¹ Translated by Rupert Precht

make predictions about the world in order to improve the quality of life of ourselves and others. Specifically, we wish to make predictions about anaesthesia, immortality, animals, teleporters and robots. What class of relation do we need for a theory capable of formulating these predictions? Here we will examine the power of each one.

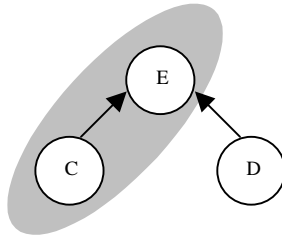
Correlation

Correlation differs from the other relations in that it is merely *observational*, while the others are *theoretical*. We simply make observations and note that two events A and B always appear together, but without making any assumptions about how or why this happens. Mere correlation, such as that of people with heat in our example, *cannot* be used to make predictions². In our heated house example, if we heated a room tomorrow, would the correlation still be maintained? Alternatively, if we placed people in a room tomorrow, would the room become heated? To make such predictions, we must go *beyond* the correlation and form a causal or identity *theory*. In this example, there are three possible theories:

- | | |
|-----------------------------|---|
| (C1) People cause heat. | (e.g. when they get cold they turn on the radiators.) |
| (C2) Heat causes people. | (e.g. the radiators may be pre-programmed, and the people ignorant of the switches, so they just follow the heat around the house.) |
| (I) People <i>are</i> heat. | (e.g. the radiators are duds, but the people give off body radiation.) |

Causation

Causation gives a theory of sufficiency but not of necessity. Knowing that C causes E tells us that wherever there is C, there is also E. But it does not provide a complete theory of E, since D may also cause E without the need for C. There may indeed be any number of additional causes of E. A causal theory always leaves open the possibility of further causes outside its *domain* (the gray area in the diagram):



In our example, knowing that the power switch causes heat is a *useful partial* theory of heat, since it allows us to make predictions about how to produce heat:

- Q: 'How can I produce heat'.
 A: 'By turning on the power switch.'

Similarly, if we found a cause of consciousness, say GABA receptor activity, we could answer:

- Q: 'How can I produce artificial consciousness?'
 A: 'By producing GABA receptor activity.'

But a causal theory $C \Rightarrow E$ (where ' \Rightarrow ' abbreviates 'causes', since the logic is identical to logical implication) is not powerful enough to answer questions of the form 'Is E caused by X?' except in the special case where X is C, e.g.

- Q: 'Can heat also be caused by coal fires?'
 A: No answer. (The theory 'power switches \Rightarrow heat' says nothing about coal fires.)
- Q: 'Can consciousness also be caused by anything other than GABA receptor activity?'
 (e.g. in aliens or robots who have no GABA receptors.)
 A: No answer.

² To do so would involve making an implicit assumption about the existence of one of the other kinds of relations.

From this it also follows that a causal theory $C \Rightarrow E$ lacks the power to answer questions of the form ‘How can we *negate* E?’. For even if we negate C, there may be D remaining, outside the domain of the theory, which still causes E. For example:

Q: ‘How can I remove heat from the house?’

A: No answer. (Outside the domain of the ‘power switch \Rightarrow heat’ theory could be other independent causes of heat, such as the coal fire.)

Q: ‘How can I anaesthetise the brain?’

A: No answer. (Outside the domain of the ‘GABA \Rightarrow consciousness’ theory could be other independent causes of consciousness, such as 40Hz oscillations.)

A causal theory of consciousness would thus answer some but not all of our questions. If we possessed such a theory, $C \Rightarrow \text{consciousness}$, we would (in theory) be able to build conscious machines, conscious copies of ourselves, and make positive confirmations of consciousness in animals which had C. But the theory would not be powerful enough to answer questions about anaesthetics or the presence of consciousness in entities lacking C.

Identity

Identity theories are the strongest of our three relations. In contrast to the open-endedness of causal theories, they provide a *closed* definition of the form ‘C is E’ (abbreviated ‘ $C=E$ ’). That is, C is E and *nothing else is*. Ontologically, identity claims are again much stronger than causal claims. The causal $C \Rightarrow E$ says nothing about what C and E are, or how they are connected. C may be merely the first of a long chain of causes, unmentioned by the theory, leading eventually to E:



...but an identity theory is committed to denying this possibility, and asserts there is just a single entity which has two names; ‘C’ and ‘E’:



We can see that identity theories are going to much harder to prove than causals, but that they are the only class with the predictive power required by our practical questions. (In addition to this practical utility, they are also more intellectually satisfying since they provide an ontological explanation.)

Note that our discussion is still at a very abstract level, and that ‘identity theory’ is not taken to mean identity with any particular class of entity. We are certainly not yet claiming that C must be physical or neural or anything else. Our choice of the identity class of theory still leaves open the possibility that the entity to be identified with consciousness could be a dualist *res cogitans*, GABA receptors, or the mind of God. All that we have selected is a *style* of explanation – we have not yet constrained the contents of that explanation. What we are hoping to do, then, is:

- (a) to say what consciousness *is*, and
- (b) to explain how other entities interact with it.

Broadly, there are two possibilities. Either:

- (1) consciousness is identical to something we already know about, or
- (2) it is identical to something we do not already know about.

In the former case we may already possess an answer for (b): for example, if consciousness is identical to 40Hz oscillations, then no extra work is required if we already know how other entities interact with 40Hz oscillations. The latter case involves postulating a new class of entity, and in this case further work will be required to formulate what Chalmers (1996) calls ‘bridging principles’ – laws which describe the interactions between old entities and the newly postulated.

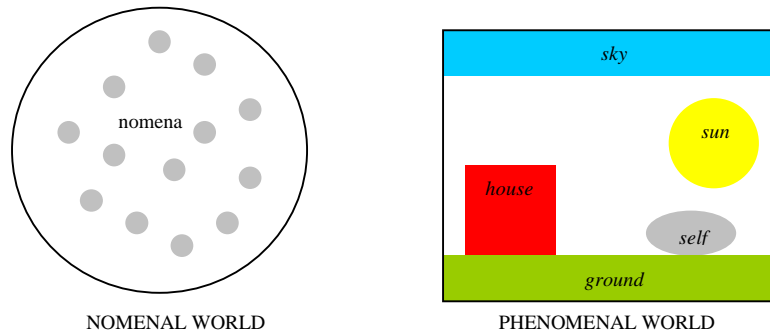
Nomenal and Phenomenal Realities

At this point I would like to formalise a claim mentioned during our earlier discussion of concepts:

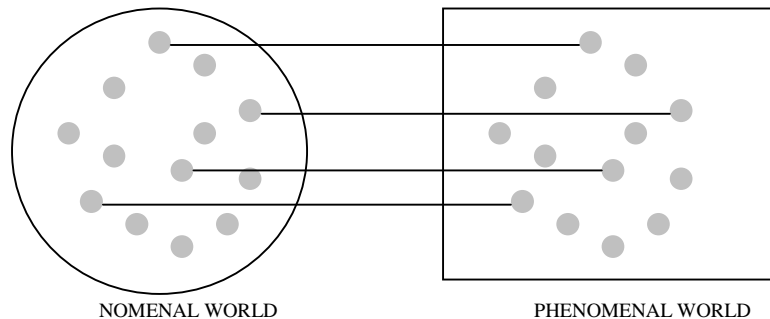
Objects exist in the phenomenal world, not in the 'real', independent-of-human-observation world. This 'real' world only contains fundamental entities: quarks, quanta, a vat, the mind of God, or whatever they may be. It is beyond the limit of our epistemology to ever know for certain what is in the 'real' world (we can only hope for our theory to converge towards it). However, assume we have magically discovered that the 'real' world is ultimately made of things called *nomena*³. Now, there is no house in the 'real' world. There is just a collection of nomena arranged into a house shape, and nothing more. The house is something we invent to describe that set of nomena in a useful practical way. Houses, suns and selves *do* exist in our phenomenologies, but not in the 'real' world. (from p.46.)

(This is a standard *Scientific Realist* position.) It is at this point that I want to retire the unqualified word 'real', (hence the careful use of inverted commas earlier) and replace it with the two more precise expressions 'nomenally real' and 'phenomenally real' – or just 'nomenal' and 'phenomenal' for short. So now, for example, the house is phenomenally real, but not nomenally real; all that exists nomenally is a bunch of nomena. The nomena exist nomenally but not phenomenally.

The nomenal world is not directly accessible to us; rather, each person has a set of concepts which are used as moulds to construct phenomenal objects from sense data, and it is only these phenomenal objects that we access as perception:



Although we can never access the contents of the nomenal world directly⁴, we can attempt to form concepts which *approximate* them. Science and philosophy progress together to provide predictions and descriptions (respectively) of the world. When science produces a new predictive system, philosophy tries to produce concepts which fit it; to infer a maximally efficient ontology. (It could be said that the mathematical predictions of science are a syntax, whilst the conceptual models built from them are a semantics). This interplay can be seen at work in contemporary quantum theory: the physicists have produced a purely mathematical prediction system, and we see philosophers hunting for interpretations such as the multiple worlds and subjective reduction stories. As the predictions get better, we assume that their associated ontologies are converging onto the structure of nomenal reality. If we could ever achieve a 100% accurate predictive model, Occam's Razor would suggest that maximally efficient phenomenal ontology fitting it would be a true description of the nomenal:



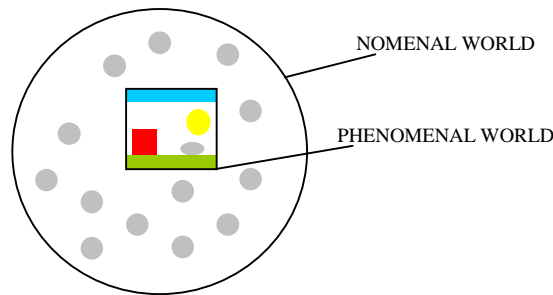
³ I have invented the word 'nomena' to be reminiscent, though not the same as, Kant's 'noumena'. My nomenal world is going to be the counterpart of my *phenomenal* world. I should make it clear that I do *not* wish to import any Kantian baggage attached to 'noumenal', hence my fresh choice of word. My nomenal world is *only* what I have defined it to be. (This footnote has been repeated just to drive the point home!)

⁴ With one exception: I will later argue that consciousness is nomenal – the only nomena which we are certain of.

Of course, we can never be 100% certain of our predictive power, since our model could always fail tomorrow and have to be updated. But each time we make a successful prediction, our confidence tends *towards* certainty. Let us call our conceptual model of the nomenal the *pseudonomenal*⁵. Note that due to the ‘seven plus or minus two’ content limitation of the phenomenal, it is unlikely that the phenomenal can ever contain the entire model at any moment. Rather, the concepts of the pseudonomenal are stored in unconscious memory, and instances of small parts are brought into phenomenal focus when needed.⁶

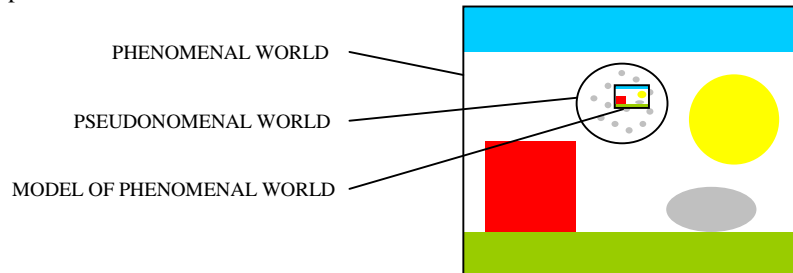
Views on the problem of consciousness

The problem of consciousness is usually conceived (by Western thought) as the problem of fitting the phenomenal world into the nomenal world - or rather, into our *model* of the nomenal world. Current scientific orthodoxy treats the nomenal as primary (and in fact often ignores the existence of the phenomenal entirely). Recall that we are looking for an identity theory: we wish to find or postulate something pseudonomenal which is identical to consciousness:



Idealist views, such as those of Berkeley, Velmans and many eastern traditions, conceive the problem the other way round: the phenomenal is taken as primary, and the problem becomes that of how to account for the nomenal world in relation to it. This view was until recently unfashionable with western scientists, but has recently begun to reappear. For example, the Copenhagen ‘lack of’ interpretation of quantum mechanics thinks of the pseudonomenal not as an approximation to a nomenal reality, but simply as a tool for predicting phenomenal experience. Honderich (2001) recently defined consciousness as “what it is for a [phenomenal] world to exist”. The primacy of the phenomenal is also echoed beautifully by one of Sachs’ (1985) neurological patients: “I had completely lost my visual field to the left, and with this as would sometimes happen, the sense that there was (or ever could have been) any *world* on the left.” (my italics).

But note that even if we are idealists, the problem of consciousness is *still there*. Recall once again that the whole point of our study of consciousness is to make practical *predictions*. We make predictions using our pseudonomenal model. If all that exists nomenally is Berkeley’s ‘mind of God’, then this model becomes a kind of predictive ‘Cognitive Theology’. Even if there is *no* nomenal world at all, and the pseudonomenal is *only* a predictive device (a position which Blackburn (1984) calls *quasi-realism*), we must *still* find a place for consciousness within the pseudonomenal in order to make our required predictions about consciousness:



⁵ I use a fresh word rather than ‘metaphysics’ in order to avoid the philosophical baggage associated with the latter.

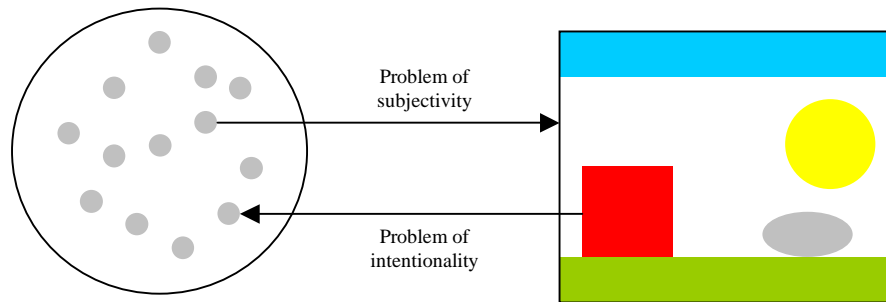
⁶ I have some sympathy for mathematical Platonism, but in an unconventional sense. We noted earlier (p.40) that there are many possible systems of mathematics, some of which happen to correspond to structures that we observe when doing physics. New systems are constantly being proposed by mathematicians, and occasionally one of these happens to be a useful tool for physics, so gradually becomes adopted into common use. Meanwhile, less-useful systems fade into disuse. This process of mathematical evolution results in the common system of mathematics tending towards the true structure of the nomenal world. The classic example of this was the invention of curved space geometry, popularised by relativity theory.

Hut and Shepard (1997) portray idealism as ‘turning the problem upside down’, and go on to introduce the notion of ‘turning the problem sideways’. On this view, the nomenal and phenomenal both exist on equal footing – neither is primary – and the problem now becomes to account for each world in terms of the other. I would like to work with this idea, as I think it can be used to frame the two classical problems from the philosophical literature (e.g. Searle, 1992):

Subjectivity: how to accommodate a phenomenal world in the nomenal world?

Intentionality: how do phenomenal objects refer to the nomena they represent?

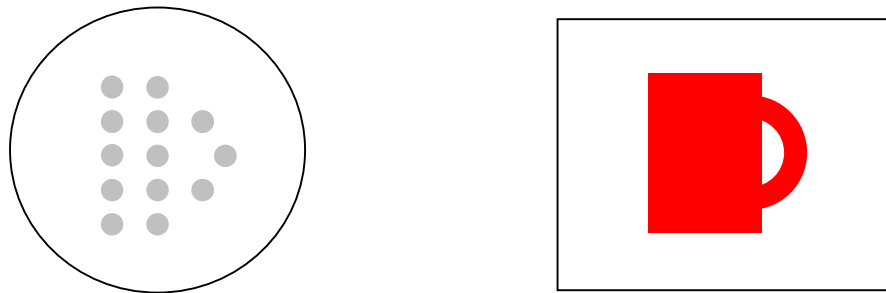
We now see that these two properties are symmetrical:



Compound objects exist only phenomenally

We now have a picture of two worlds, nomenal and phenomenal. Within the phenomenal, we form a pseudonomenal model of the nomenal, which converges upon the nomenal with the progress of science and philosophy.

The nomenal world is the set of all entities which are *fundamental*. For example, cups do not nomenally exist. We can perceive a cup as a single phenomenal object, but we know that nomenally, there is nothing more to the cup than its parts. If we think, for example, that quarks are fundamental, then we accept that were are nomenally just looking at a bunch of quarks – there is no additional cup entity in the nomenal world. The cup object exists only phenomenally:



Anything which is not fundamental does not exist nomenally. For, like the cup, if there is a way of breaking it into subcomponents, then we see that nothing more than the subcomponents is needed in the nomenal world.⁷

Objection: Those who balk at the claim that cups do not nomenally exist may attempt a *translation* strategy to rescue them. The word ‘cup’ is to be used in two different languages: a phenomenal language and nomenal language. In phenomenalese, cup_p refers to the unitary cup object. In nomenalese, cup_n refers to the collection of nomena which comprise the cup:

$$\begin{aligned}\text{cup}_p &= \text{cup object} \\ \text{cup}_n &= \{\text{nomenon}_1, \text{nomenon}_2, \text{nomenon}_3 \dots \text{nomenon}_n\}\end{aligned}$$

⁷ Margaret Thatcher’s infamous comment ‘There is no such thing as society’ is often negatively misinterpreted: what she really meant was that there is nothing *more* to society than the individuals who comprise it, in the same way that there is nothing more to the cup than the nomena which comprise it. She would have been clearer (to philosophers) if she had said ‘Society exists phenomenally but not nomenally.’ (She is showing herself to be a socio-individual identity theorist rather than a society emergentist – ideas which we will discuss shortly.)

However, this is merely a linguistic shorthand. The translationist has produced a *different* definition of ‘cup’ (i.e. cup_n) and has defended the existence of *his* definadum. I have no objection to the existence of this set of nomina. But when *I* use the word ‘cup’ (or any other phenomenal object name, of course) in this thesis, I mean it in the sense of cup_p , and my claim, that cups do not nomenally exist, still holds.

Rival pseudonomenologies and cultural relativism

It is interesting to note that different people view the same nomenal world in different ways, due to their differing conceptual frameworks. These frameworks are like different pairs of glasses, each providing a different view of the world.

As we discussed in our model of concept formation, the way that we form our most basic concepts is by generalisation. We receive sense-data⁸ from ten red objects, extract the statistical similarity between them, and form the concept RED. Instances of RED then can enter our phenomenology when we are exposed to red objects. We tend to form such concepts to correspond to the things that we see most often in the world, since they give the strongest statistical patterns. (Recall the generalisation operator, λ , of our model.) To a tribesman living in the rainforest, concepts of FIRE, WOMAN, and DANGEROUS_THING would perhaps be amongst the first to form. We are also *taught* concepts by others – but usually these have also been selected (by the teacher) for their usefulness to us in our environment.

Having concepts which correspond to statistical patterns in our sense data means that they have predictive value (assuming that the statistical patterns persist in the future). Scientific experimentation is essentially the process of deliberately and frequently exposing ourselves to environments whose behaviour we wish to predict more accurately. By creating this exposure, we form and refine concepts which capture and so predict regularities in those environments. This is pseudonomenology in action.

This *phenomenal* reality of predictive useful objects is what Dennett speaks of in his discussion of ‘Real Patterns’ (1991b). Dennett defines ‘real’ objects as those whose postulation gives predictive power. Reality is to be taken as a continuous variable, with an objects degree of reality equal to its predictive utility. Centres of Gravity may be very real as they are useful and accurate; tree spirits are less real for the opposite reason. His ‘reality’ is our pseudonomenality: these useful and predictive objects exist inside our phenomenal world. (Which is ironic, since Dennett elsewhere (1991) denies the existence of the phenomenal world and everything in it!)

So, exposing ourselves to different environments (including deliberate experimental ones) tends to produce different pseudonomenologies; that is, different sets of concepts through which we phenomenally perceive the nomenal world. This is why different cultures have different beliefs, even though they may all be practising science (science being the *process* of systematically experimenting to refine pseudonomenology; not any particular set of results). For a tribal culture, conceiving of the weather as the activities of a set of Gods may be a successful result of science, *if* the model gives useful predictions about the weather. The oriental model of *chi* (a kind of energy which flows through the body) gives predictions about how to perform acupuncture which are currently unexplainable by our western view of the body. *Feng Sui* postulates spirit-like entities which flow through buildings – and gives useful predictions about interior design which are currently beyond western architectural theorizing. Western science, on the other hand, has focussed its attention on mechanical environments, and has produced models to successfully predict the behaviour of the bewildering array of technology which we rely upon today, and which is beyond the reach of the non-western models mentioned above. Freud’s attention was focused almost entirely on mad, middle-class women, so his model gives good predictions about them, but perhaps not for other sectors of society.

Imagine a high-technology society, like ours, but which has evolved in a somewhat bizarre environment: it lives under the cover of tall trees whose leaves each have tiny double-slits; and the sun above the forest only emits one photon at a time. The reader familiar with quantum theory will recognise this set-up as a version of a famous experiment which shows Newtonian Physics to give incorrect predictions. In our own society, this kind of environment is somewhat obscure, so our pseudonomenology was only updated to take account of it at a relatively late stage in our development.

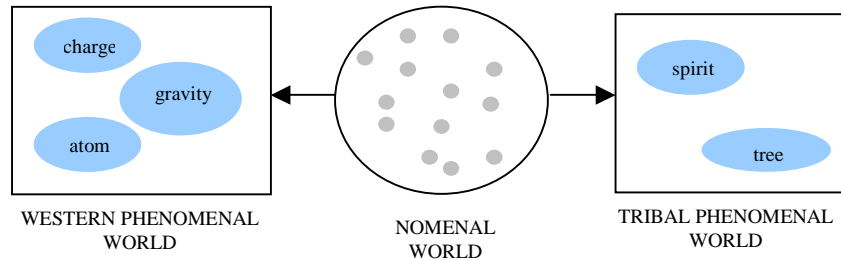
⁸ This is not to say that the sense-data reaches consciousness.

However, the tree-slit society encounters it regularly, so they may well never form a Newtonian pseudonomenology but instead go straight to quantum-like models at an early stage.

The point of all this is to show that *Feng Sui* spirits, Newtonian particles, quanta and *chi* are all just pseudonomenal models of the nomenal world. *None of them exist nomenally* – though we hope that by directing our experimental attention to a broad range of environments⁹, we will be able to converge our model onto the structure of the nomenal world.

A brief anti-Cognitivist diversion

It is interesting to take a brief diversion here to remove a pretension of Cognitive Science. Our picture, is one of a nomenal world perceived in different ways by people with different conceptual ‘glasses’:



The relationship of a phenomenal view to the nomenal world is the same as that of a high-level computer program to its compiled binary code. Imagine two programs, (P1 and P2), in different languages, which happen to compile down to the same binary code (BC) – that is, they both have the same *function*:

```
(P1):  { for I=0 to 10; print "hello"; next I }
(P2):  { I=0;   while I<10 {print "hello"; I=I+1;} }
(BC):  010001011101010100000010110100101001010010100001
```

Now suppose we are given BC, but not told where it came from. We are shown P1 and P2 and asked: ‘Which is the *real* high-level structure of BC?’. Clearly, we have no way of telling which program was used to compile into BC – since they are equally good high-level descriptions of its function.

The point is that neither P1 or P2 exist *within* BC – they are just conceptual frameworks for understanding and predicting BC’s behavior. In fact, BC might not have been compiled from any high-level program at all – it may be an evolved string of code produced by a genetic algorithm. In this case, the question of which of program is the ‘correct’ description becomes completely meaningless.

If we think of the brain as an evolved, low-level computer, we now see the analogy with Cognitive Science, which seeks to describe its workings. Long-standing debates such as that over ‘single vs. dual route mechanisms of verb inflexion’ (Seidenberg & McClelland, 1989; Colthart *et al.*, 1993) are arguments over which of two (or more) models describes what is ‘really’ going on in the brain. Is there ‘really’ a single PDP route for verb inflexion or are there ‘really’ two symbolic routes? Both opponents are producing more and more refined models of their claims to account for the psychological data. We can imagine a situation where both models are so refined as to match the data exactly – as P1 and P2 match the behavior of BC. But which would be correct? Is there one route or two? These are non-questions. Nomenally, the routes do not exist. All that exists is a bunch of nomena. The entities in the models exist only phenomenally, as predictive models, like *chi* and *Feng Sui* spirits.

Physics as pseudonomenology

There is an important difference between Physics and the other sciences: Physics is the study of *fundamental* entities. No-one would claim that the entities postulated by *Feng Sui* (e.g. spirits), Biology (e.g. cells) or Cognitive Science (e.g. inflexion routes) exist nomenally (other than in the

⁹ There is an intriguing question here as to how we should go about selecting environments for this ‘broad range’. Modern science tends to direct its attention towards environments which are useful to, and fundable by, humans; but is this really the best way of converging upon nomenality?

translationist sense discussed and dismissed earlier). Those entities are useful phenomenal fictions for making predictions. But Physics seeks a greater goal: it seeks to construct a model whose structure mirrors that of the nomenal world itself - the pseudonomenology.¹⁰

It has scientific and philosophical components which construct predictions and pseudonomenal concept models respectively. I define the nomenal world as the set of all fundamental entities which exist. The goal of Physics is to achieve convergence of the pseudonomenal onto the nomenal; to accurately and completely describe the nomenal world.

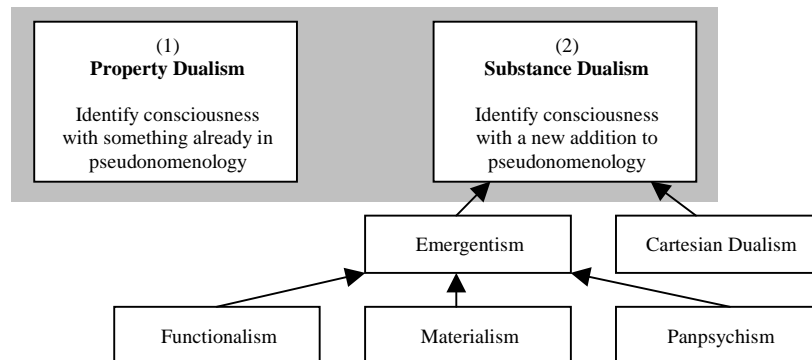
Consciousness is nomenal

Unlike cells and *Feng Sui* spirits, **consciousness nomenally exists**. It is not just a phenomenal predictive fiction. Like Descartes, I know that it exists non-fictitiously because I *am* it. I am not a fiction! (Furthermore, even if I was a fiction, consciousness could still not be phenomenal, because phenomena exist *inside* consciousness, and an entity logically cannot be identical to something inside itself.) It is therefore a nomenon – the *only* nomenon that we know directly¹¹.

We wish to locate consciousness in (our model of) the nomenal world. There is something in the nomenal world which is identical to consciousness, and the goal of our science is to find that identity. As the study of all fundamental entities, Physics is incomplete if it fails to place the consciousness nomena into its pseudonomenal model.

Reducing the ‘isms’

Armed with this identity characterisation of the problem of consciousness, I will now analyse some major contemporary philosophical positions on the place of consciousness in nature. I will show that *all* of them reduce to substance dualism. Recall that broadly, there are two classes of solutions to the identity problem: either (1) consciousness is identical to something already in our pseudonomenology, or (2) it is identical to something yet to be postulated by it. The former solution is property dualism – we add new properties to an existing entity. The latter is substance dualism – we add a whole new entity. My argument takes the form of a chain of reductions of one theory to another. It may be useful to sketch the battleplan in advance, where the arrows mean ‘reduces to’:



Cartesian Dualism: two worlds

Descartes' position was that consciousness is 'non-physical', by virtue of existing outside time and space. Instead, it exists in a different 'domain', *res cogitans*.

¹⁰ Some objectors would say that Physics is only concerned with making predictions, not concepts, the latter being the task of philosophy. However, as discussed in the introduction, the predictions are presented *as* concepts, e.g. the concept of mass contains predictions about the typical behavior of mass. Predictions and concepts are inseparable. Some physics journals have banned papers concerning interpretations of quantum theory, claiming that is a philosophical matter. But all physicists *do* use interpretation concept models when thinking about quantum theory. To deny this is reminiscent of the behaviorist denial of consciousness: everyone thinks about it but no-one is allowed to speak of it.

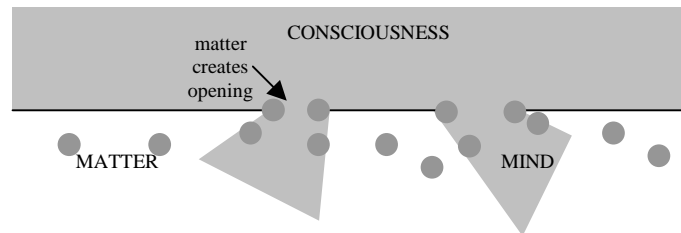
¹¹ It may be objected that I have committed the fallacy of excluded middle by claiming that consciousness is not phenomenal so is therefore nomenal. But the objector would then have to postulate an entirely new class of entity which neither exists nomenally nor phenomenally. The onus would then be on him to justify this, since Occam's razor is on my side.

However, we have defined physics as the study of all things which are fundamental, not just those which exist in space and time. We have not made any claims about what nomena are or in what domain they must exist. If Descartes is right and consciousness exists outside space and time, then we must simply extend our physics to encompass *res cogitans*, since physics must study all that is fundamental. (There is nothing sacred about confining physics to spacetime - contemporary physicists already try to *derive* spacetime from more fundamental entities such as 10-dimensional ‘superstrings’.) So Cartesian Dualism simply reduces to solution (2): postulate some new pseudonomenal entities and identify consciousness with one of them.

Emergentism

There is currently much talk about consciousness as an *emergent* property. I think that emergence is an incoherent concept, but I will try to recount its story before disposing of it. The story is that consciousness is something ‘more than the sum of its parts’; it is ‘irreducible’ but still ‘arises from’ the parts. Imagine we build a radio receiver from pieces of metal, turn it on, and hear Bach. The Bach is not reducible to the properties of the metal – it is something more. Emergentism treats consciousness like the Bach – the brain is a necessary condition for consciousness to ‘arise’, but consciousness is not identical to the parts, it is something more.

Actually, this idea is not new. Schweitzer (1993) reviews the pseudonomenology of Sāṅkya-Yoga philosophy, dating from around the 4th century BC. Like contemporary Cognitive Science, Sāṅkya-Yoga draws a divide between Mind and Consciousness: Mind is merely the non-subjective, functional organisation of matter; Consciousness is the non-functional, subjective experience accompanying Mind¹². Sāṅkya-Yoga places all consciousnesses outside space (like Descartes), but treats them as a *single* entity: all people are to be conscious from the same source. Consciousness is like a sun shining above the clouds. When matter is arranged in a particular configuration (Mind), this creates a channel, a gap in the clouds, for consciousness to shine through. Mind is matter illuminated by consciousness:



Contemporary Emergentism is essentially just the same story. Matter, arranged in a certain configuration (like a radio receiver), provides the condition for consciousness (like Bach) to appear in that matter, but without being a property of that matter alone.

Emergentism is often framed as follows: Let N be the set of all fundamental entities (the nomenal world). A subset, S , of N is arranged in a certain way, and this causes a emergent entity E to arise which is not already in S . S is defined to be the complete cause of E .

This is incoherent. For if E is something more than S , then something additional to S exists, call this X (for ‘extra’). Now, either X is fundamental or isn’t. In the latter case, X thus reduces to a collection of fundamental entities rather than a single one. In either case, there is a set of fundamental entities X_N which ‘arise’ as an effect of S . Because they are fundamental, they are already in N , by definition of N . But because they also (by definition) play a role in E , and because S is defined as the subset of N which is the complete cause of E , then X_N must also be a subset of S . But now all of E , including the alleged ‘something extra’, is in S . And this contradicts the claim that something extra has ‘arisen’. We see that the alleged emergent entity was, by definitions, already *in* the set of things which caused it.

A less technical version of the above argument can be formed using the Sāṅkya-Yoga example. The emergence claim is now that S the collection of fundamental entities, presumed to be matter, which is the complete cause of consciousness. When S is arranged in a certain way (Mind), consciousness arises from it. But consciousness is not determined entirely by S , it is something extra. The ‘light rays’

¹² The Indian text uses different words, of course, but they translate loosely into ‘Mind’ and ‘Consciousness’.

on the diagram are the X_N – the extra part of consciousness Mind not determined by S. Here we see the obvious error. The conscious mind was *not* caused entirely by the presence of matter, but also by the presence of the dualist world of consciousness. So S, the complete cause of consciousness, did not just consist of matter after all, but also of dualist substance. The supposed emergent entity, consciousness, was already *in* the set of things from which it was supposed to arise!

Thus we see that *Emergentism is in fact just a mildly incoherent form of substance dualism*. Just like Sāṅkya-Yoga, it is the claim that matter, when arranged in a certain mystical way, opens the gates for an immaterial entity to flow in. Physics is the study of all fundamental entities, so would again be extended with pseudonominal concepts to model the new immaterial entity. Once these entities are added to our ontology, it is easy to see how the Emergentists made the mistake of leaving them out from the set of entities causing consciousness to ‘arise’.

Functionalism

(Philosophical) Functionalism is the doctrine that consciousness is identical to a type of functional state, regardless of its implementation. It implies that we can implement consciousness equally well using microchips, neurons or cogs. A common formulation is that there are different kinds and degrees of consciousness, so that *all* functional systems possess some kind of consciousness. This is an elegant idea: a thermostat is a simple system, so has a simple consciousness; anthills, humans, cities, and corporations¹³ are complex systems, so possess a higher consciousness; the entire universe is a huge complex system, so is the most conscious entity in existence, God. Sadly though, and despite its elegance, functionalism is incoherent. We will show that the only way to rescue it is transform it into an emergence model – which then reduces to substance dualism as shown above.

First, though, I would like to defend functionalism from a poor class of ‘arguments from author’s sense of absurdity’ (as discussed in the introduction). I do this to distinguish my own, better, argument from these weak ones. Examples of these arguments include Searle’s Chinese Room (1980) and Block’s ‘Nation of China’ (1978). A huge functional system is postulated, with one or more humans as its tiny components¹⁴. The system is arranged so as to produce the same functional behavior as a single (conscious) human. The author of the argument then states that he thinks it is absurd that this system could be conscious, and concludes that therefore the system is not conscious and therefore functionalism is wrong. These arguments are not valid because the inference ‘I think X is absurd \Rightarrow X is false’ is not a valid inference. Aristotelian physics thought it ‘absurd’ that ‘a ball dropped from the top of a moving ship’s mast will land at the base of the mast’, but that statement is still true. The whole point of functionalism is that it *postulates* that functional systems are consciousness – functionalists do not think that this is ‘absurd’. The whole debate then degenerates into a brawl of ‘I think X is absurd therefore it is false!’ versus ‘I think X is obvious therefore it is true!’ from which I think it best to walk away.

As we discussed at the start of this thesis, the business of philosophy¹⁵ is to *define* concepts and questions, not try to answer them. Trying to answer them non-empirically generally leads to embarrassment, and the ‘absurdity’ debate is a prime example of this in action. Answering the question ‘Is functionalism true’ is not something that can be done philosophically. However, what we *can* do is to *un-ask* the question: we can show that it is ill-formed and invalid. That is how I propose to

¹³ It would be wonderful if corporations were conscious, since it would provide me with a way of causing pain to my bank when they provide bad service. This was the case in the old days when bank managers maintained personal, moral, relationships with their customers. But now corporations are impersonal entities, we can no longer treat them as members of our moral community and are forced to take an ‘objective stance’ toward them. (cf. Strawson, 1962). All I can do now is close my account and move to another bank, which is much less efficient for both parties than moral punishment leading to an improvement of the service.

¹⁴ Searle’s Chinese Room is deliberately misleading with regard to the size of system required – implicitly leading the reader to assume that the system consists of one human and a few books. In fact, to be capable of carrying out a realtime conversation in Chinese, the amount of books required would probably be such as to require an oil-rig-sized library, and the human would to work at something approaching the speed of light to respond in time. When this is pointed out, Searle’s ‘intuition pump’ is revealed – it is not the human but the whole system, of which the human is a *tiny* part, that is the candidate for having consciousness. (For a beautifully pictorial version of this exposé, see Hofstadter, 1997). Searle’s reply, that the human could internalise the contents of the oil rig and still somehow perform the calculations in real time, is simply a logically impossible assumption, given our present concept of how the brain works; and having assumed a contradiction it comes as little surprise that he then able to prove anything he wants.

¹⁵ At least for the purposes of the present thesis.

dismantle functionalism – by showing it to be an incoherent concept. My argument is similar to Schweitzer's (1996) devastating critique of functionalism, but I have recast it in terms of my concept theory, as I will later build on it to prove something more powerful.

As far as our best current pseudonomenology is concerned, *functions do not nomenally exist*. Imagine a machine built of cogs and moving parts. The behavior of the parts is governed only by the laws of physics. Now it happens that this machine has been designed, by a conscious human, to calculate sums of numbers. The numbers are 'input' as the positions of two cogs, C1 and C2, and the result 'output' as the position of a third cog, O1. Now, the input, output, and hence function, of the machine are all objects in the human's phenomenal world – he perceives the parts *as* a machine. We can demonstrate this by the following humorous possibility: suppose another man has, by a remarkable chance, invented the same machine, but as a means of computing something completely different – say, weather patterns in China. He perceives ten different cogs, C3 to C12, to be the inputs, and ten more cogs, O2 to O11, as the outputs (showing wind speed, direction, rainfall etc.). Now recall the anti-cognitivist arguments from earlier – the situation is analogous. Again, the two men will argue over what is the 'real' function of the machine, and again, there is no answer to the question. The machine just 'is what it is' – a bunch of nomena – and the inventors are just perceiving it in different ways. Nomenally, there are no functions, only nomena.

The situation becomes even sillier when, with Honderich (2001), we replace the discrete set of cogs with the less discrete human body. Honderich imagines his own body as a supposed functional machine, and again pictures two people arguing over his functional description. The first wishes to treat only Honderich's ears and mouth as inputs and outputs (perhaps to conduct a Turing Test). The second treats his entire biology including weight, sweating, hair growth and excretion as outputs; and his food intake and respiration as inputs. Which of them has the 'real' functional description? Functions exist only in the phenomenal world of the beholder¹⁶.

Consciousness exists nomenally. I know that because I exist nomenally, not fictionally. Functions and machines are fictions; they do not exist nomenally. (Unless we wish to reconstrue our entire pseudonomenology to accommodate functions as fundamental, which would be an unworkable prospect due to the multiple-perceptions argument above.) One cannot make an identity between something that nomenally exists and something which does not. So therefore the notion of functionalism – the hypothesis that consciousness could be identical to function – is incoherent.

One way to attempt a rescue of functionalism is to rebuild it as an emergence theory: rather than the functions themselves being identical to consciousness, the presence and arrangement of the underlying nomena, which cause functions to be perceived, *also* cause consciousness to emerge. The function is demoted back to a *correlate* of consciousness. This is like saying that the arrangement and activity of nomena in a radio receiver cause both the perception of that receiver and the emergence of music. So functionalism becomes Emergentism, which in turn reduces to substance dualism as we have already seen.

Materialism

Materialism is the thesis that conscious experience is identical to some macroscopic object. For example: identical to a complete brain state, or to 40Hz oscillations, or to GABA receptors. Ryle (1950) gives the classic analogy of consciousness to Oxbridge universities: there is nothing more to the university than the colleges which constitute them¹⁷; and nothing more to consciousness than its brain parts. Ryle says that to think there must be something more is to make a 'category mistake'.

Now I hope that it will be reasonably obvious from the above discussion of functionalism that materialism is also incoherent, for the same reasons. At risk of repeating myself to the point of irritation, macroscopic objects *do not exist nomenally*. There is no University of Cambridge in the *nomenal* world; there are no γ -oscillations or GABA receptors or brain states. All of those things exist only phenomenally – *inside* consciousness. Consciousness exists nomenally. And one cannot make an identity between a nomenally existent thing and a nomenally nonexistent thing. Ryle's analogy is

¹⁶ This is a pun: the phenomenal world is, of course, the 'I'. (As opposed to the self-object in it, which is the 'Me'.)

¹⁷ Most American tourists have not read Ryle, so great fun can still be had in misdirecting them to Anglia Polytechnic when they ask 'Where is the university?' whilst standing right opposite King's College.




flawed because it identifies a mere *phenomenal* entity (the university) with another set of mere phenomenal entities (the colleges). Ironically, then, it is Ryle himself who has made a ‘category mistake’ by confusing nomenal and phenomenal categories in his analogy.

Materialism is today championed by Churchland (e.g. 1996), in the form of ‘Eliminative’ materialism. The idea is that neuroscience is currently an immature science, and has not yet established the ‘correct’ set of concepts to carve up the parts of the brain. It may be incoherent to identify consciousness with any of the current parts, but we should ‘wait and see’ how neuroscience progresses before addressing the identity question. Churchland is still adamant, however, that neuroscience is the right level to look for the consciousness identity¹⁸, and specifically that sub-neural objects will be irrelevant. Of course we can see that simply reorienting our macroscopic neuroscientific concepts in this way will still never provide a coherent identity – since such identities would still be doomed attempts to identify a nomenal entity with a non-nomenal entity.

As for functionalism, we could rescue materialism by rebuilding it as an emergence theory: the presence of the nomena, arranged in ways which causes us to perceive the correlating objects, also causes consciousness to emerge. But emergence reduces to substance dualism, so this theory could hardly retain its name ‘materialist’!

It is worth noting that Chalmers’ (1996) theory that consciousness is identical to *information* may also fall foul of the same arguments. Under the classical¹⁹ definition of information – ‘a difference that makes a difference’ – different people can perceive information in different ways. For example, consider the first figure below:



The figure contains some information and some noise. Perhaps it is a visual message that has been sent down a noisy channel. Now, how much information is in the figure? A western human would probably say that the pixels corresponding to those in the ‘a’ (shown in the second figure) carry the information, and the pixels corresponding to those in the ‘’ (the third figure) are information-less noise. But what if I was to tell you that ‘’ is actually the 56th letter of a rare tribal alphabet? To a tribesman, ‘’ is the signal and ‘a’ is the noise. As there are more letters in the tribal alphabet than in ours, the original image is perceived as carrying more information by the tribesman than by a westerner. Thus we see that (classical) information is merely a phenomenal construct and hence cannot be coherently identified with nomenal consciousness.

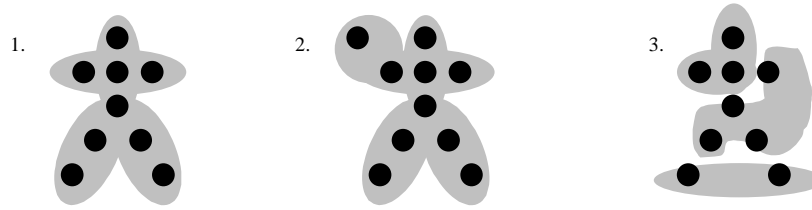
Panpsychism

Panpsychism is the theory that *all* nomena have a some ‘simple’ form of consciousness, and that these simple forms combine to create more complex consciousness. So a single electron is a bit conscious, a molecule is a little more conscious, a blood cell more still, a human very conscious, and possibly cities, corporations and God even more so. Versions of panpsychism have been formulated by Spinoza (1677), Leibniz (1695) and, more recently, by Whitehead (1933).

There are two classic objections to panpsychism, both of which are ‘arguments from the author’s sense of absurdity’. These are the ‘problems’ of *combination* and *permutation*, and are best illustrated pictorially:

¹⁸ Perhaps this is because she is a neuroscientist?

¹⁹ There is an alternative formulation of Chalmers’ theory which does not fall foul of these arguments, which we will examine later.



(1) shows a human consciousness – constructed from the parts of the body. (2) is the combination problem: a single dust particle (say), taken together with the body, forms another system – and panpsychism would have to grant a *separate* consciousness to that system – and to all other systems made up of the body plus anything else! (3) is the permutation problem: within the body *itself* there are an almost infinite number of subsystems, each of which must be granted its own consciousness. The arguments then state that these consequences are ‘absurd’ and therefore that they must be false. As with the anti-functionalist arguments earlier, this does not disprove the (admittedly unintuitive) proposed hypothesis – it merely shows it to be unintuitive.

However, once again we can instead attack the theory by showing it to be incomplete or incoherent. Panpsychism postulates that a set of fundamental nomena, S, ‘combine’ to produce a new nomenal entity, consciousness. Consciousness is the grey ‘stuff’ in the above diagrams – something *additional* to the black parts. Now, if consciousness was *already in* S, then the theory is incoherent – since S cannot combine to produce itself. Panpsychism would not be required – since consciousness was already a dualist substance, existing independently of the rest of S. Alternatively, if consciousness was *not in* S, then we are left with an emergence theory: something nomenal has ‘arisen’ which is more than its parts. Panpsychism thus reduces to Emergentism, and thus to substance dualism as we have already seen.

What is in the nomenal world?

We have seen that consciousness is a nomenon²⁰ which we wish to locate in our phenomenal model of the nomenal world. This location is to be performed by identifying it either with something already in our model (‘property dualism’) or with something to be newly postulated (‘substance dualism’). We have exposed several ‘isms’ which made the mistake of identifying consciousness with non-nomenal entities. But we have not yet said much about what entities the nomenal world (or rather, our best model of it) *does* contain. We noted earlier that contemporary physics tends to avoid the question of what *entities* to postulate, preferring just to provide mathematical predictions. For most scientific applications, this predictive behavior is sufficient. But a science of consciousness is different: because of the nomenality of consciousness, the content of the pseudonomenal world now becomes of vital importance, as it drastically constrains our identity theory.

We need to examine our current conception of the pseudonomenal. If we find a pseudonomenal entity which *already correlates* with consciousness, then Occam’s Razor suggests that the most likely model of consciousness in nomenality is that which identifies it with that correlate. If we do not find a pseudonomenal correlate, then we will have to postulate the new identifying entity. As preparation, let us first examine some historical pseudonomenological formulations.

A brief guide to pseudonomenologies

Cartesian Dualism was born of Newtonian Physics. The pseudonomenology of Descartes and Newton was a world of three-dimensional Euclidean space, populated by spherical objects, each bearing the properties of mass and charge, and evolving deterministically through time according to a small set of laws. The ontology thus comprised: space, time, spherical objects, mass, charge, fields and laws. Now, do any of those entities obviously correlate with consciousness? No. Hence we resort to the less parsimonious Cartesian option: consciousness is not identical with any of those ‘physical’ things, but is identical to a new entity, postulated specially for the purpose. (Additional pseudonomena, namely new fundamental laws to connect the consciousness entities to the ‘physical’ entities, are also required.)

²⁰ Meaning that *each* consciousness is *a* nomenon. Compare the grammar with ‘*the* electron is a nomenon’. Perhaps due to its unusual history, the word ‘consciousness’ has rather odd grammatical properties.

Do there *nomenally* exist both ‘objects’ and ‘properties’ – and if so, what is the difference between them? Most so-called properties are just phenomenal objects. For example, an atom’s property of ‘being hydrogen’: *nomenally*, all that exists are the subcomponents of the atom, arranged in a certain way. It is the conscious observer who perceives them as a hydrogen atom. Similarly, John’s property of ‘being tall’ is just a phenomenal perception of the arrangement of the *nomena* comprising John, and nothing more. So could it be that there are no *nomenal* properties, only phenomenal perceptions of *nomena*? No. For if *nomena* had no properties, then they would all be identical; and there would exist nothing but a single *nomenal* unity from which nothing more could be constructed! *Nomenal* objects, *if they exist*, must therefore have *nomenal* properties. A neat way to distinguish *nomenal* properties from phenomenal ones is to model *nomenality* as a high-dimensional spacetime, with dimensions for each of the properties. Relativity already conceives of mass as singularity in spacetime; string theory replaces other properties with dimensions. This picture of *nomenality* thus dispenses with objects and properties – its ontology consists *only* of high-dimensional spacetime. Could anything in this picture be identified with consciousness? Possibly one of the dimensions, though there is no immediately obvious correlation.

It is interesting to note here that our western-scientific angst over objects and properties does not exist in some other cultures. Our language is built around nouns and verbs, building the object/property dichotomy into our thought. ‘X is a triangle’ and ‘X is triangular’ mean the same thing, yet we are tempted to think of the triangle as object and property respectively, when in fact there is no difference. The triangle is just phenomenal object when we intend the meaning of the sentence. In comparison, some Indian languages are *verb*-based: from an idealist viewpoint, they describe the dynamics of experience rather than postulating pseudonomenologies, rather like Copenhagen quantum theory. A stronger, *nomenal* realist interpretation of those languages would construe pseudonomenology of consisting of *processes* rather than *objects*. This is a very strange idea for us to comprehend – but we should remember that as we have no direct access to *nomenality*, anything is possible there! Whitehead (1932) is one of the few in our tradition to hold a similar view; though it is currently becoming fashionable consider Heidegger’s metaphysics of ‘becoming’ and ‘being’ (TSC conference, private conversations) which appear to resemble its spirit.

I give the above examples to try to break down our prejudice that *nomena* must resemble macroscopic objects (the ‘little balls’ picture). Once this prejudice is overcome, the gates are opened for all kinds of alternative formulations of *nomenality*. Trope theory (e.g. Campbell, 1990) postulates entities from which both objects and properties are to be derived; whilst Wheeler (1990, cited by Chalmers, 1996) has reformulated quantum theory treating *information* as pseudonomenal. This formulation of information is not the same as the mere phenomenal information we examined earlier – it is supposed to exist independently of the observer. (We now see how Chalmers can coherently propose a *pseudonomenal* identity between consciousness and information which escapes the earlier criticism.)

Consciousness is quantum state reduction

However, it is not necessary to resort to reformulating our entire pseudonomenology in order to find a pseudonomenal correlate – since we *already* have a candidate in our contemporary view! Recall our earlier discussion of superposition and Quantum theory. Quantum theory requires the addition of an extra *nomenal* component to our high-dimensional spacetime ontology: state reduction, **R**. We have already seen a theory – Penrose-Hameroff – which proposes a subneural mechanism for superposing parts of the brain, and whose reduction is to correlate with moments of consciousness. Now, regardless of whether the implementation details are correct or not, this is the *only* type of correlation we have examined which is postulated to exist *pseudonomenally*, as so is our *only* current candidate²¹ for a coherent identity. As Stapp pointed out, all quantum interpretations *already contain* a subjective element involved in the reduction process. There is no need to postulate new dualist substances or reformulate pseudonomenology because we already have a ready-made pseudonomenological correlate to identify with. *If*, as Penrose and Hameroff suggest, we can demonstrate the correlation empirically, (recall that **Orch-OR** requires very special circumstances to occur, and the brain’s microtubule

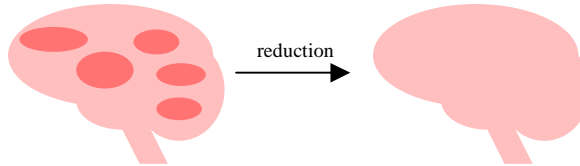
²¹ The theory is extremely speculative and controversial, even with the Consciousness Studies community. However, it is nevertheless our only current pseudonomenal candidate. I challenge anyone who wishes to dismiss it non-empirically to come up with a better suggestion for a *pseudonomenal* correlate. (If you can’t find one then you must default back to substance dualism, which is surely even less plausible than the P&H correlate.)

environment is the only such place we know of in the universe), then we have found our maximally parsimonious identity.

R-Identity solves the Problem of Unity

Identifying consciousness with state reduction also solves a problem which even traditional substance and property dualisms fail. The problem is that consciousness is both unitary and compound at the same time. Why is this? It is *unitary* because it exists nomenally. If it was simply a *collection* of nomena, then it itself – the collection – would not be nomenal, because collections are abstract objects which exist only phenomenally. We could claim that the elements of the collection shared a property which ‘bound’ them together, but then we are left with the panpsychist problem: what is this binding but a new, unitary nomenal entity? But consciousness is also *compound* because it contains structure – the objects of phenomenology which we so painstakingly classified earlier in this thesis. *Prima facie* it appears that there is a contradiction here and that consciousness is an incoherent concept. But this would be to deny one’s own existence, which is also a contradiction!

However, the apparent problem of unity is based on an outdated Newtonian-Cartesian pseudonomenology, consisting of ‘little balls’ in spacetime. In this picture, it is true that no nomena can be compound. But the addition of the new (pseudo)nomena, state reduction, to contemporary pseudonomenology solves the problem. State reduction is not a ‘little ball’ – it is an *operation* on (potentially large) volumes of spacetime. Suppose a large volume of brain (shown shaded below) becomes superposed and reduces:



The reduction operates on a volume, which contains other pseudonomena. It is a pseudonomenon which ‘contains’ other pseudonomena in a pseudonomenally real way. (Contrast this with the claim that γ -oscillations solve this same ‘binding problem’: γ -oscillations are not pseudonomenal, state reduction is.) In this way, the abstract information contained in the superposed volume becomes transferred to the phenomenal world, but without being *identical* to it (cf. Chalmers). It is the reduction itself that is the identifier.

Note that the superposed volume can be *distributed* over many non-connected locations; binding and reduction still work the same way even if the separate parts are miles apart²². A likely way that Penrose Reduction could produce our phenomenology structures would be as follows: each concept is stored by a trained neuron or assembly of neurons, which also functions as an instance detector for that concept. A moment corresponds to reduction of a set of superposed concept-detectors, each of which produces its corresponding object in the phenomenology. A phenomenal object is thus the reduction of its neural concept detector. The shaded areas in the diagram above might correspond to HOUSE, SELF, SUN, RED, YELLOW and LINE concept detectors; with the color and shape detectors located in visual cortex, and the higher-level concepts located further forward.

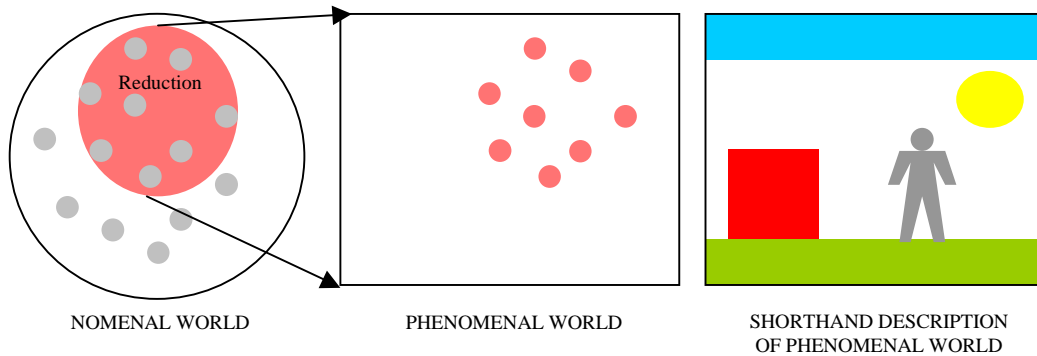
The Problem of Multiple Unities

There is a further, related problem which I call the Problem of Multiple Unities. *Within* the phenomenal, there are unitary objects – e.g. houses, trees, redness and self. Reduction accounts for the binding of brain nomena into *one* unity – but how can we account for the existence of these unitary subcomponents?

A first (and flawed) solution attempt might go as follows: there *are no* unitary objects in phenomenology – the objects are our *shorthand* for describing the collection of millions of nomena

²² Reduction of distributed superpositions has been suggested as a method for transmitting information faster than light: two objects are superposed together in London, then one is separated off and taken to Edinburgh. The London party then collapses their object, which *instantaneously* causes the Edinburgh object to collapse also. (Private conversation with Imperial College undergraduate.)

which become transferred into the phenomenology by the reduction. The phenomenal does not really consist of houses and trees, but of tiny parts:



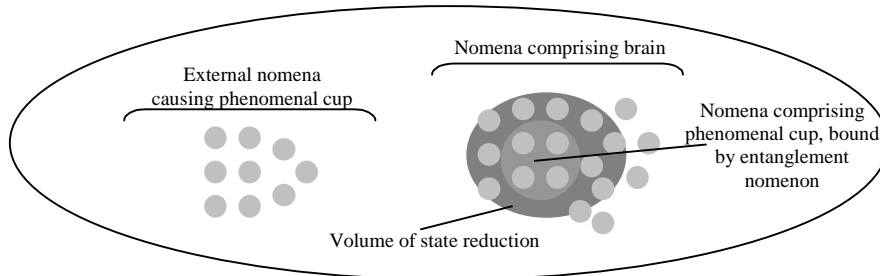
However, this is not how I experience my phenomenology, so it is wrong. I *do* experience objects as unities.

A refined version of this story might claim that the number of nomena which actually reduce is tiny - say only seven plus or minus two - and that each of them corresponds to a single phenomenal object. Each concept assembly could contain just one reducing nomenon at its core. However, this would be incompatible with the Penrose gravitational reduction, which requires a much more substantial amount of mass to be involved in the reduction.

So we are still left with the problem of how a large number of nomena become bound *within* phenomenology; state reduction only answers the question of binding the *whole* phenomenology. I admit that my knowledge of physics trails off at this point, but I note that as Physics provided us with a solution for the whole-binding problem, so it might also provide a solution for this sub-binding problem. I would like to point vaguely towards the recent notion of *entanglement* in quantum theory, which appears to allow spatially distributed pseudonomena to be bound in multiple ways within a superposition, as a possible solution²³. This is not meant as a full scientific solution to the problem – just as an illustration that it is *possible* for Physics to find solution to the problem. I am not yet qualified to explore this suggestion further, but I would like to throw it to Physics for further discussion.

Phenomena are also nomenal

An interesting consequence of the Multiple Unity problem is that phenomenal objects actually turn out to be nomenal! – though *not* in the obvious way. When we have a phenomenal cup object in our phenomenal world, it *nomenally* exists as an entanglement (or whatever) *inside* our brain. In contrast, the nomena *outside* which cause the perception are still just unbound nomena; it is the *internal*, phenomenal cup which is a nomenon:



It is tempting to call these phenomenal nomena ‘representations’, since they are used in the brain to guide its behavior regarding the external nomena which causes them. *However*, I dispute this word usage. These phenomena *are* my phenomenal world – they do not *represent* it. I live in a phenomenal world: when I see a cup in front of me, it is a phenomenal cup. This cup in my brain is not a

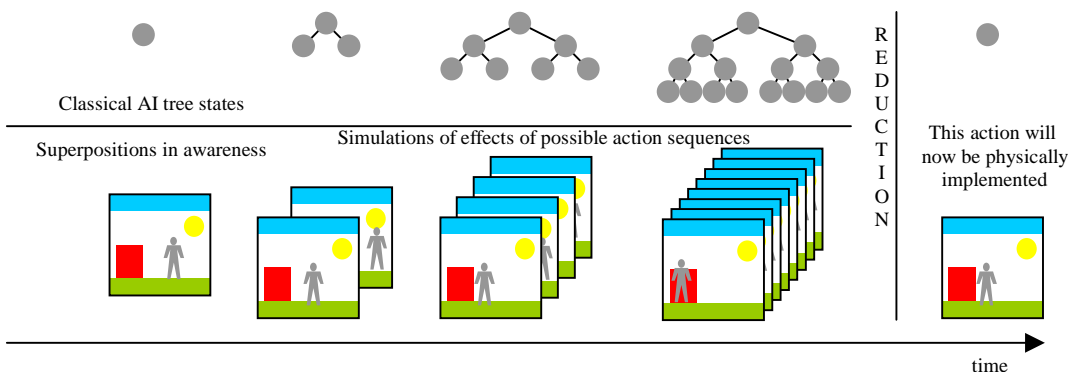
²³ Joseph Chen (University of Hamburg), private conversation.

representation of that cup; it *is* that cup! (Note that there is no contradiction with phenomena also being nomenal: indeed, the aim of our science is to place the phenomenal into the nomenal.)

Once we have allowed the phenomenal to become nomenal, an obvious question to ask is: why do we need ‘consciousness’ as an additional nomenon? Why couldn’t consciousness just be identical to a collection of such phenomenal nomena? The reply to this is that it is a form of panpsychism: it postulates nothing but a collection of protoconscious nomena, with no means of binding them into unitary consciousnesses. As we showed earlier (p.70), panpsychism must then postulate a new entity to perform this binding – which brings us right back to postulating consciousness as a separate nomenon.

State reduction and free will

A further advantage of the state reduction identity is that it provides a neat explanation of how classical determinism is compatible with conscious free will. The content of awareness²⁴ routinely becomes superposed, into a multitude of possible future worlds which simulate the consequences of the subject’s possible future actions. The reduction collapses this multitude into a single world, whose physical presence in the brain may then cause the appropriate action to be implemented. It is easy to imagine classic AI tree-searching taking place here: with the thickness of the superposition (i.e. the number of superposed possibilities) growing as the tree’s depth is searched. (Recall that the reduction is supposed to occur at around 50Hz, so each moment’s decision will probably be very simple.) In the diagram below, the man is working out how to get to the house:



This kind of parallel searching is precisely what quantum computers are good at (hence their use in RSA cryptography cracking – see Nielsen & Chuang, 2000). Consciousness is thus identical to the process which selects which future phenomenal world to enter. It thus has quantum-causal power, but without interfering with any laws of current physics - which simply model this ‘free will’ choice as random. This is as good an explanation of free will as we could wish for!²⁵

Making the identity: from prediction to explanation

We have begun bridging the gap between first and third person consciousness data, and have seen how to move from the observation of correlations to the making of predictions by means of postulating an identity that is parsimonious with our conception of the pseudonomenal world. For Science, this seems to be as far as we need go, since the purpose of Science is to generate useful predictions, and that is what we can now do.

However, some philosophers may demand more: they call for an *explanation* of our identity postulation. The literature is filled with discussions of the so-called *explanatory gap* (e.g. Levine, 1983; McGinn, 1989) – the notion that reaching such an explanation is forever beyond our abilities. ‘Gappists’ hold that there is a form of explanation which can be given for ‘normal’ identities, and which is impossible to give for the phenomeno-physical identity. My purpose in this section is to show

²⁴ Remember that the contents of awareness only becomes phenomenal *during* the reduction, not before it.

²⁵ This picture of consciousness navigating through multiple phenomenologies is reminiscent of Schopenhauer’s *The World as Will and Idea* (1819).

that gappists are wrong and that ultimately the forms of explanation of both consciousness and ‘normal’ identities are of equal power.

What is to be identified

We will proceed by examining the power and form of ‘normal’ identity explanations, using the classic ‘water is H₂O’ statement as our example. Suppose that Science has found a correlation between the presence of water and H₂O, and is successfully postulated an identity which gives useful predictions. But the gappists claim there further exists an explanation of *why* they are identical, which goes beyond the correlation and predictive power.

It is important that we first clarify exactly what is meant by the words ‘water’ and ‘H₂O’ – obviously, if the words initially have the same *meaning* then the identity is trivial. Instead, the words are each supposed to stand for concepts which have been formed from collections other concepts. So ‘water’ stands for the concept WATER, which has been formed from concepts such as WET, DRINKABLE, LIQUID, RAIN. Recall from our concept theory that we can form the WATER concept either by being taught this definition explicitly, or by generalising it from a collection of exemplars. So we might begin with:

$$\text{WATER} \equiv \lambda(\text{stuff_in_this_glass}, \text{stuff_in_this_swimming_pool}, \text{stuff_falling_from_sky})$$

where each exemplar is a particular perception of a collection of lower-level objects. Over time (and probably taking place during sleep), the regularities are extracted and the definition becomes a mix of processed higher-level concepts and a few less-processed exemplars:

$$\text{WATER} \equiv +(\text{LIQUID}, \text{WET}, \lambda(\text{stuff_in_this_swimming_pool}, \text{stuff_falling_from_sky}))$$

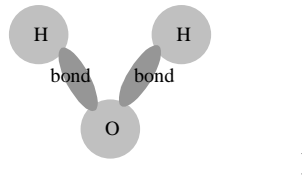
And eventually it may become fully abstracted:

$$\text{WATER} \equiv +(\text{LIQUID}, \text{WET}, \text{DRINKABLE}, \text{RAIN})$$

The concept H₂O is formed by a similar process, but from chemical, scientific perceptions:

$$\text{H}_2\text{O} \equiv +(\text{HYDROGEN}, \text{HYDROGEN}, \text{OXYGEN}, \text{BOND})$$

(Recall that ‘+’ is a shorthand for a set of spatial connectivity relations – the H₂O concept really looks something like



So, the type of identity that the gappists think they can ‘explain’ is that of WATER= H₂O, which consists of two concepts that initially have different definitions in terms of subconcepts.

‘Explaining’ Identities

Now that we have clarified the question, here is how the gappist’s ‘explanation’ proceeds²⁶. He processes the set of subconcepts of H₂O to show that *they* are identical to the subconcepts of WATER. The form of the explanation is thus:

$$\begin{aligned} \text{WATER} &\equiv +(\text{LIQUID}, \text{WET}, \text{DRINKABLE}, \text{RAIN}) \\ \text{H}_2\text{O} &\equiv +(\text{HYDROGEN}, \text{HYDROGEN}, \text{OXYGEN}, \text{BOND}) \\ \underline{+(\text{LIQUID}, \text{WET}, \text{DRINKABLE}, \text{RAIN}) \text{ is } +(\text{HYDROGEN}, \text{HYDROGEN}, \text{OXYGEN}, \text{BOND})} \\ &\text{WATER is H}_2\text{O} \end{aligned}$$

²⁶ Our discussion of Marr’s visual concept recognition theory (p.49) was a prelude to the following argument.

Or for gappist identity explanations in general:

$$\begin{array}{l} A \equiv \{P\} \\ B \equiv \{Q\} \\ \{P\} \text{ is } \{Q\} \\ \hline A \text{ is } B \end{array}$$

Note that once we have made such an identity, we are able to *merge* the two identified concepts into a single concept, containing both sets of properties. This is often what happens during ‘Eureka!’ moments: we suddenly realise that the concept VOLUME_OF_WATER_DISPLACED_FROM_BATH can be merged with VOLUME_OF_BODY_PLACED_IN_BATH; or that SIEGMUND’S_FATHER can be merged with SIEGLINDA’S_FATHER (in *Die Walküre*). The discovery of mathematical equations is another example: we realise that two mathematical concepts are equal, i.e. identical. In our H₂O example, we may form a single new concept, C, which identifies (and so replaces) both of the old ones:

$$C \equiv +(\text{LIQUID, WET, DRINKABLE, RAIN, HYDROGEN, OXYGEN, BOND})$$

The gappist claims that there is no explanation of this type for the phenomeno-physical identity. The latter identity, he says, relies only on correlation and predictive power observations, but ‘normal’ identities such as WATER=H₂O can be *explained*. Therefore, he continues, we should not be as confident of the phenomeno-physical identity as of ‘normal’ identities. However, I will argue that the gappist’s identity explanation amounts to *nothing more than* a collection of correlative-predictive observations of exactly the same kind that our phenomeno-physical identity is based upon; and hence that the latter is just as strong as any ‘normal’ identity.

My first question to the gappist: how do you explain that ‘{P} is {Q}’? Simply assuming this would be no better than assuming ‘A is B’ – it simply shifts the ‘explanatory gap’ onto a different identity.

In response, the gappist will attempt to show that each subconcept in Q can be deduced from P together with background knowledge; and that each subconcept in P can be deduced from Q together with background knowledge. For example, water’s property (a.k.a. subconcept) of ‘being the stuff that rain is made of’ is to be deduced from the properties of H₂O. The gappist will say something like “H₂O is the stuff that rain is made of, because H₂O is made of hydrogen and oxygen, and hydrogen and oxygen bound together are found in clouds and fall from these clouds under gravity.”:

$$\begin{array}{l} \text{RAIN} \equiv +(\text{FALLS_FROM_CLOUDS, IS_IN_CLOUDS}) \\ +(\text{HYDROGEN, OXYGEN, BOND}) \Rightarrow \text{IS_IN_CLOUDS} \\ +(\text{HYDROGEN, OXYGEN}) \ \& \ \text{GRAVITY} \Rightarrow \text{FALLS_FROM_CLOUDS} \\ \hline +(\text{HYDROGEN, OXYGEN, BOND}) \Rightarrow \text{RAIN} \end{array}$$

The gappist will provide a set of such ‘explanations’ to account for all the required deductions of subconcepts. Many people would want to stop here, simply accepting this ‘explanation’. But like the small child constantly asking ‘why?’ we may continue. *Why* do hydrogen and oxygen fall from clouds? *Why* does:

$$+(\text{HYDROGEN, OXYGEN}) \ \& \ \text{GRAVITY} \Rightarrow \text{FALLS_FROM_CLOUDS}$$

Again, to assert this as a brute fact is no better than just stating that water is H₂O. The gappist promised us an *explanation* and still he is relying on a brute assertion. We may as well have asserted that water is H₂O and saved ourselves the effort.

So now the tired gappist proceeds to tell us yet another layer of his ‘explanation’, perhaps: “hydrogen and oxygen fall under gravity, because they have mass, and all masses fall under gravity.”

$$\begin{array}{l} \text{HYDROGEN} \quad \equiv +(\text{MASS, ...}) \\ \text{OXYGEN} \quad \equiv +(\text{MASS, ...}) \\ \text{MASS} \ \& \ \text{GRAVITY} \Rightarrow \text{FALL} \\ \hline \text{HYDROGEN} \Rightarrow \text{FALL} \\ \text{OXYGEN} \Rightarrow \text{FALL} \end{array}$$

Again, the gappist has expanded the explandum into yet more subconcepts, and asserted a brute fact (a law of gravity) to perform the ‘explanation’. Again we see that his ‘explanation’, when *expanded* into now hundreds of similar low-level ‘explanations’, amounts to *nothing more* than a bunch of brute assertions.

Does the gappist wish to concede defeat at this point, and acknowledge that his ‘explanation’ consists only of *brute assertions*? A set of brute assertions is even less convincing an ‘explanation’ than our phenomeno-physical identity that the gappist is supposedly attacking: at least *we* have correlation and predictive power evidence to back us up! But proceeding beyond this point would involve explaining *why* mass falls under gravity, which is quite possibly beyond the capability of the gappist. And even if he scraped through that task by telling us some story about gravitons or relativity, he would do so by producing a new explanation of exactly the same form as the last, which would leave him with an even harder brute fact to explain, and so on, *ad infinitum*.

But, perhaps in desperation, he protests: “But the law of gravity is an empirical law of nature! It’s not just the *brute assertion* that you accused me of! It is an *empirical law* which has been formulated from many careful observations of correlations between mass, gravity and motion; it gives predictive power! That’s much more convincing than just brute assertions!”

And so, my dear gappist friend, your so-called ‘explanation’ of identity amounts to *nothing more* than a set of observed correlations and predictive utilities. Which gives my phenomeno-physical identity *precisely the same power* as your ‘explanation’. If there is an ‘explanatory gap’ between phenomenology and physics, it is only as wide as the ‘gap’ between water and H₂O.

The concept of consciousness

Now that we have shown that there is no problem in identifying the concept of CONSCIOUSNESS with some correlating concept, say QUANTUM_REDUCTION, it is interesting to ask exactly what the new, merged concept contains. Most people probably form QUANTUM_REDUCTION by some analogy to macroscopic objects – typically taught with the superposed cat example. They imagine two cats appearing from a single cat – then imagine an observer looking at them and the disappearance of one cat. This is then analogised down to the quantum level. (And refined with further definitions and examples if the person makes a serious study of physics. The layman’s concept probably goes little further than the cat analogy.) Penrose’s reduction uses a further analogy: spacetime as a 2D sheet which bubbles and bursts:

$$\begin{aligned}\text{QUANTUM_REDUCTION} &\equiv (\text{CAT1}, \text{CAT2}, \text{DISAPPEARANCE_OF_CAT1}) \leftrightarrow (\text{MASS1}, \text{MASS2}, ?) \\ \text{PENROSE_REDUCTION} &\equiv (\text{SHEET}, \text{BUBBLE}, \text{BURST}) \leftrightarrow (\text{SPACETIME}, \text{SUPERPOSITION}, ?)\end{aligned}$$

At the start of this thesis (p.8) I gave an informal definition of consciousness:

I define ‘my consciousness’ as that which is common to the experiences that I have when I am awake or dreaming, and which I do not have when I am in dreamless sleep, under anaesthesia, or unborn. I assume that other humans, including the reader, also have similar feelings and so are able to understand this definition. Once one possesses the concept ‘my consciousness’, one may then draw an analogy between oneself and others in order to obtain the concepts of their consciousnesses. By generalising over all these concepts of peoples’ consciousnesses, I define the concept ‘consciousness’, and invite the reader to do the same.

We can now use our concept theory to semi-formalise this:

$$\begin{aligned}\text{MY_CONSCIOUSNESS} &\equiv \lambda(\text{all conscious experiences}) \\ \text{OTHER_CONSCIOUSNESS} &\equiv (\text{SELF}, \text{MY_CONSCIOUSNESS}) \leftrightarrow (\text{OTHER}, ?) \\ \text{CONSCIOUSNESS} &\equiv \lambda(\text{all other_consciousnesses of everyone we have conceived})\end{aligned}$$

Recall that our generalisation definitions tend to become feature-extracted into composition definitions. It is interesting to ask how MY_CONSCIOUSNESS might be feature-extracted in this way: what is it that is common to all our conscious experiences? As I remarked in the introduction (borrowing from Honderich and Revonsuo):

Roughly, the feeling that I have generalised in my definition of consciousness is one of the existence of a three-dimensional world, with my self in it.

We could notate this as: $MY_CONSCIOUSNESS \equiv +(MY_WORLD, SELF)$

After studying Buddhist philosophy, we might be convinced that the SELF is no longer part of the definition, so we might redefine as: $MY_CONSCIOUSNESS \equiv MY_WORLD$

So ultimately, our definition of CONSCIOUSNESS is cast in terms of experience of a WORLD; projected onto others by analogy to our self; and generalised over those others. We cannot know what the worlds of others are *like*, so are unable to extract features from the latter generalisation – it must remain in that form.

We can see that there are no subconcepts shared between the concepts of consciousness and reduction. We can also see how no subconcepts would *ever* be shared between the subjective definition of consciousness and any objective scientific concept: since consciousness is defined ultimately in terms of MY_WORLD and my SELF, and these are precisely the concepts which are ‘banned’ in objective definitions (by definition of ‘objective’!).

So McGinn is right that we can never produce any layers of gappist ‘explanation’ of the sort discussed earlier. We have to make do with a single layer of predictive correlation observation explanation. But as we showed above, this form of explanation is what *all* ‘explanations’ ultimately reduce to anyway, so is no problem. We are free to simply merge together the concept of consciousness with the concept of our best *pseudonomenal* correlate – which currently happens to be quantum state reduction:

$$NEW_CONSCIOUSNESS \equiv +(CONSCIOUSNESS, QUANTUM_REDUCTION)$$

A verificatory gap

The gappist argued that any phenomeno-physical identity is weaker than physical-physical identities because the latter are backed up by an *explanatory* power unavailable to the former. We have shown this to be false. But there is another reason why the former could be held less justifiable than the latter: non-consciousness identities can be *verified* with a method unavailable to consciousness identities. Here I demonstrate this asymmetry and discuss whether or not it is a problem.

Recall the distinction between *causal* and *identity* theories. An identity theory $C=E$ gives us predictive power for questions of the form, ‘Does E result from C?’, ‘Does E result from X?’ and ‘How can I remove E?’. But a causal theory $C \Rightarrow E$ can *only* answer the first class of question. We chose to formulate an identity theory of consciousness so as to achieve answers for all of the questions.

An additional relation to identity and causation (a.k.a. sufficiency) is *necessity*. ‘C is necessary for E’ will make predictions about the third question (how to remove E), but not about the other two. Denoting this relation ‘ $C \Leftarrow E$ ’, here is a table summarising which relations can answer which questions:

Prediction types	Theory types		
	$C=E$ (Identity)	$C \Rightarrow E$ (Causation)	$C \Leftarrow E$ (Necessity)
$C \rightarrow E$	✓	✓	
$X \rightarrow E$	✓		
$\neg C \rightarrow \neg E$	✓		✓

Now, the potential problem is that for non-consciousness identities, predictions of all three forms can be *verified* by experiment. We can show that such theories are ‘predictively correct’. But for consciousness identities, we can *only* demonstrate necessity. We cannot demonstrate that the whole *identity* is ‘predictively correct’.

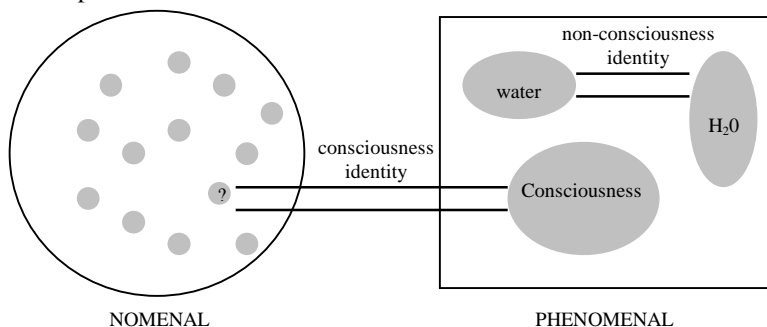
This is because we cannot observe the presence of consciousness from the third person. One can only detect consciousness that one *is*. We can verify the necessity prediction ‘ $\neg C \rightarrow \neg \text{consciousness}$ ’ by removing C from our *own* brain and observing our own loss of consciousness²⁷. But to verify ‘ $C \rightarrow E$ ’, or the more general ‘ $X \rightarrow E$ ’ would require an experiment where C or X is *isolated*, and the presence or non-presence of E observed. For non-consciousness E this is no problem: we produce C in a laboratory

²⁷ Obviously not a direct observation, but by inferring it from the apparent passage of time while we were unconscious.

and look for E. But when E is consciousness we are unable to make this third-person observation. We are incapable of stripping down our own brains to contain just C, and our brains might not contain X, so we are equally unable to perform the experiment in the first person.

So, unlike non-consciousness identities, the consciousness identity can only be verified up to necessity. We can never verify it *as an identity*.

The verificatory gap is of course a symptom of the fact that non-consciousness identities are always between *phenomenal* objects, whereas the consciousness identity is between a phenomenon (some pseudonomenon) and a nomenon (consciousness); and the fact that we can observe phenomena but not nomena in the third person:



The best we can do is to make the identity of consciousness with a pseudonomenal concept. But just as we can never verify that our pseudonomenology is correct, so we can never show that our identity *with* a pseudonomenon is correct either.²⁸

Is the verificatory gap really a problem for the strength of our identity? This is a difficult question, but I think that the answer is no, for the following reason: We should ask, 'Why did we choose to postulate the identity?' Was it to make those particular predictions? No: we postulated it because we knew that consciousness must be identical with *something*, and in the presence of a perfect correlate, the most *parsimonious pseudonomenology* is obtained by identifying it with that correlate. The justification for the identity was parsimony, not prediction verification. And once we have justifiably made the identity, we get the predictions for free. The predictions were never part of the justification.

A parallel could be drawn with other identities, for example that of water and H₂O. The *justification* for the identity is given, ultimately, by a set of correlations (eventually obtained by expanding the 'explanation', as discussed earlier). *After* the identity has been made, it provides us with a potentially infinite set of predictions ('water will behave like H₂O in situation X, for all X'). But we would not expect to see *every* prediction of the identity to be verified before we accept it (that would be impossible!). The justification was given by the *correlations* (a.k.a. the 'explanation'); the additional predictions then come for free.

Certainty *versus* confidence

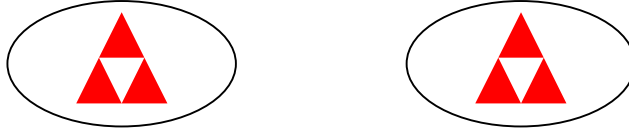
We can never be entirely certain of the truth of the non-necessity predictions of our identity theory. But if we accept that *theory* with confidence C, then we should hold the same level of confidence in those predictions even though we cannot verify them. C depends on the resolution (i.e. the level of detail) of the correlation (recall the cover picture), and its ability to stand up to the best falsification tests we can throw at it. We should ask, 'What is the likelihood of this correlate occurring by chance?' If we manage to find a pseudonomenal *creature* correlate occurring over the exact areas of the brain which bear the *transitive* correlate of phenomenology, this probability becomes negligible, and C becomes close to certainty. So we should then accept the identity. The situation is similar to an example which I mockingly call 'the hard problem of stereo vision'. Suppose that your two eyes are presented with the following images:

²⁸ Another way of thinking about this is that if we *could* verify that consciousness was identical to a particular pseudonomenon, then we would have verified that that pseudonomenon is a *true description* of a nomenon, because we know that consciousness is nomenal. But this is impossible because we know that we can never be sure that our pseudonomenology is correct.

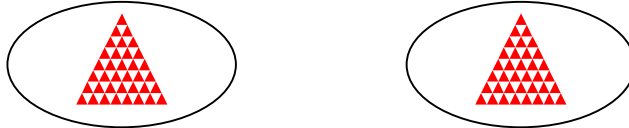
EXPLANATION



Now, there is no way of being *certain* that the triangles seen by the two eyes are identical. It may occur by chance that two different triangles happen to be presented at the same time. However, if the match becomes more detailed:



...the correlation is much less likely to have occurred by chance – you are more confident of the identity. And if you could see a correlation right down at the quantum level of detail, the probability of correlation due to chance becomes negligible:



The triangle figures above were of course chosen to be reminiscent of our cover picture. When correlation occurs at higher and higher resolutions, the likelihood of identity tends to certainty. And an exact correlate at quantum level, *if* we can find it, would put the identity beyond all reasonable doubt. When we see the same detailed structure from two viewpoints, we should simply assume that they are views of the same thing:



Sky Castle, by M.C. Escher (woodcut, 1928)

Conclusion

A sketch of a theory

“Now is the time for *doing*.”

- Tony Blair

The aim of this thesis was to work towards a useful, predictive theory of consciousness. Science and philosophy together make progress by constantly defining and refining concepts, whose definitions include expectations of future behavior. We wished to find such a concept for consciousness.

Logical findings

All concepts are grounded in our experience, and capture generalisations about patterns in that experience. In everyday life, we form concepts of macroscopic objects such as houses, trombones and coffee cups. (Recall that objects are phenomenal instances of concepts.)

One has good reason to believe that in addition to one’s phenomenal world, there exists an external nomenal world. Somehow one’s phenomenal world, and those of others, are part of this nomenal world. This is the most parsimonious story about why there are other people in one’s phenomenal world whose behavior closely resembles one’s own.

One may never access the external nomenal world directly, but through experiments and observation one is able to form a converging conceptual model of it within one’s phenomenal world. The concepts of this model can also be shared with others through teaching and learning. We called this model the *pseudonomenal* world.

The nomenal world consists of all entities which are *fundamental*. Composite objects are simply phenomenal instances of concepts which we form in order to make predictions about the massed behavior of those entities. They are predictive fictions. (If we arrange three bricks in a triangle shape, there is no nomenal triangle which comes into existence in addition to what was there before. The triangle object is just a phenomenon.)

My phenomenal world exists. Like Descartes, I know that because I *am* my phenomenal world. I am not a fiction; not a mere phenomena. (And how could a phenomenal *world* be identical to something existing only *inside* itself?) My phenomenal world (and presumably everyone’s) must exist *nomenally*. My own consciousness is a nomenon that I *can* experience directly¹. Here is the argument condensed into sequent form:

The nomenal world is the set of all entities that are fundamental
All non-fundamental objects are phenomenal fictions
<u>My phenomenal world is not a phenomenal fiction</u>
Therefore, my phenomenal world is a nomenon

So *logically*, the only way that we can place consciousness into our conceptual model is by *identifying it with a pseudonomenon*. Given our existing pseudonomenology, we have two choices: either (1) identify consciousness with an already-postulated pseudonomenon; or (2) postulate a new pseudonomenon especially for the purpose (this may also involve adding and removing other

¹ It is not an external nomenon. It is not external to the phenomenal world because it *is* the phenomenal world!

pseudonomena from the model to keep it coherent). The former strategy is sometimes called ‘property-dualism’; the latter is ‘substance dualism’. There are no other options.

Consciousness *logically cannot* be identical to anything macroscopic or abstract, such as books, cups, brains, brain states, functions, information or behavior, *unless* we readjust our entire pseudonomenology to accommodate the proposed object as fundamental.

If we can find something *already* in our pseudonomenology is a perfect creature correlate of consciousness, then parsimony suggests identity beyond all reasonable doubt. Seeing similar complex objects in each of our eyes, we simply assume that the objects are identical; the probability of it happening by chance is negligible. If we cannot find a pseudonomenal correlate we are *then* forced to default back to substance dualism and/or a radical reappraisal of our pseudonomenology.

We have shown that the so-called ‘explanatory gap’ involved in justifying our identity is only as wide as similar ‘gaps’ in *all* other scientific identities, so it not a problem for our identity. We further discussed and disarmed a potential ‘verificatory gap’: our consciousness identity cannot be *verified* to the same level as other identities; but this is not a problem because our reasons for making the identity were never based upon those verifications.

Empirical findings

Descending from the analytical to the empirical, let us now examine our contemporary pseudonomenology. Unlike in Descartes’ time, we *do* currently possess a pseudonomenon - *quantum state reduction* - which appears to be a prime candidate for consciousness identity. Penrose’s theory of quantum gravity allows reduction to occur locally, this enabling it to act *on* other nomena. These nomena could bear the *information structure* of the phenomenal world, whilst not being identical to that world: they are the *transitive* correlate of consciousness. The reduction itself could be the *creature* correlate, and hence, by the above argument, the *identity* of consciousness. There are no other correlation candidates in contemporary pseudonomenology. If quantum reduction is *not* the identity, then radical pseudonomenology reformulation is the only other option.

Thus, to save our pseudonomenology, is currently worth investing our best efforts to look for this quantum reduction correlate. We need somehow to find the phenomenological information embedded in the objective brain, and to demonstrate the reduction process occurring over it. If we can find this, then the problem of locating consciousness in our pseudonomenology is solved!

We are a long way from finding a quantum correlate. We saw that quantum coherence *can* be sustained in the brain over short distances (p.27), though this was only shown in an artificially-induced situation. Hameroff and Penrose have suggested that microtubules are a possible site for quantum effects – this seems to be the best location suggestion so far.

A research strategy proposal

However, simply suggesting locations is not enough: we need to demonstrate that the correlation is there. The higher-resolution transitive correlate we can find, the more confident we can be about the identity of consciousness with the creature correlate acting over it. To find both correlates, we need to develop formal descriptions of both pseudonomenal *and* phenomenal structures. The former task is currently a major focus of physics; the latter needs urgently to catch up.

This thesis has examined several fragments of attempts to develop a Phenomenology, and has tried to knit them together into a neo-cognitive model. The main claim of this model is that the phenomenal world consists *only* of *objects*, which are instances of *concepts*. We have removed the Humean distinction between ‘impressions’ and ‘ideas’: when we see ‘redness’, we are seeing an instance of the low-level *concept* of redness; in much the same way as seeing ‘triangleness’ is seeing a triangle object; or seeing ‘John’s-face-ness’ is seeing an instance of the concept of John’s face; or seeing ‘ π -ness’ is seeing an instance of the concept of π . This picture is backed up by neuroscientific evidence about low-level vision: it consists of a series of layers of neurons, each of which detects a feature set of a higher level of abstraction from the previous one. But the mechanism of each layer is essentially the same – there is no obvious split in mechanisms which corresponds to the split between ‘impressions’ and ‘ideas’. Husserlian Phenomenology and recent change blindness experiments provide further evidence for the theory.

In addition to the phenomenal aspect of our concept theory, we also discussed how concepts come to be formed: by generalisation, feature extraction, explicit definition and analogy. We also drew from James' research on the Fringe, which we have translated into the set of 'pointers' from one concept to another, which allow associated objects to be brought in and out of awareness. (Awareness being the functional correlate of the phenomenal world.)

The concept/object theory proposed here is still under-specified, but could form the *basis* of an implementable model of the structure of phenomenal worlds. (This is *not* to say that such simulations would be conscious!). This model, together with a suitable graphical interface, would provide a formal framework for subjects to compose phenomenological reports – something that has always been badly lacking in Phenomenology. Once these reports are input as data, the (formal) *information* which they contain can then be used in statistical processes to find and test potential objective correlates.

I am not an experimentalist, so I have not specified any particular protocol for reporting with the model (other than making the 'scene flash' suggestion). Rather, I would like to see the model implemented and handed to the experimentalists for them to 'play with' and devise their own protocols. Similarly, I am aware that the model is currently based on some very controversial and under-verified assumptions (e.g. from the Husserlian and Buddhist traditions), and I would like to see these assumptions tested more rigorously, and the model refined in accordance. My model is not intended to be some dogmatic assertion about what phenomenal worlds are like – rather it is an initial suggestion, an attempt to get something off the ground, to be passed to the Consciousness Science community for criticism and development.

How should we go about looking for the objective transitive correlate? Where should we look? What should we measure? A first, 'brute force' strategy would be to cover the subject's head with every kind of measuring device we can lay our hands on – fMRI, EEGs, MEGs, coherence recorders etc. – and simply measure *everything*. For each phenomenal report datum that the subject makes, we store the complete objective data set for the same moment that was reported. Get subjects to perform hundreds, thousands, millions of tasks, and for each one, store the two data sets. Then feed the collection of these pairs into a massive neural network, or other pattern detector, to automatically pull out the correlations. Unfortunately, this strategy would be subject to the 'curse of dimensionality' and we would need an unfeasibly large number of data to perform it².

A better strategy, then, is to work 'top-down': begin with large-scale *creature*-correlate approximations, and use them to progressively refine the search for the transitive correlate. For example, we know that γ -oscillations (colloquially known as '40Hz oscillations') are an approximate creature correlate. So perhaps the transitive structure is embedded in the Fourier transform of these oscillations? If not, then look for a better creature approximation: what is the underlying cause of the γ -oscillations? Perhaps NMDA receptors are involved in producing them, and perhaps we then discover them to be a more accurate approximation to the creature correlate. Now we can check NMDA activity patterns for the transitive correlate. If this fails, we then ask what mechanisms are involved in NMDA activity, make and test new hypotheses about which of *those* could be a new creature approximation, and check *those* for transitive correlation... Hopefully this process will lead right down to the quantum level, and to the correct brain location of the transitive correlate, where we will find the information from the phenomenal reports displayed before us in the brain!

Once we have found the transitive correlate it should be relatively easy to look for and find the *creature* correlate acting over it. If this creature correlate is pseudonomenal then we can make our identity. If it is not, then we should again search for underlying causes of this correlate until we find a pseudonomenal one. If we cannot find one, then we will be forced reformulate our pseudonomenology as discussed earlier.

Penrose and Hameroff have already jumped this gun to some extent, by placing their 'microtubule' hypothesis down at the quantum level without the need for my 'homing-in' programme. But their hypothesis does not tell us exactly *where* to look to find the vital transitive correlate: *which* neurons should we look inside? *Which* tubules? *Which* tubulins? Recall that phenomenal worlds appear to

² Anthony Jack (University College, London), private conversation.

hold only a tiny amount of information at any particular moment – finding this information in the tubulins would be like finding the needle in the haystack. There are about 10^{12} neurons in the brain, 10^7 microtubules in each neuron, and at least 10^2 tubulins in each microtubule. We cannot feasibly record data from every tubulin in the brain! I suggest that the top-down method described above may still help to guide the search towards the particular tubulins that matter.

Speculative findings: the best theory for *now*

Even before a correlation is established, quantum reduction is *still* the best identity to make for consciousness *given* our current state of knowledge. We do not yet possess conclusive evidence either for or against the possibility of there being a reduction correlate in the brain (although we are beginning to see some healthy debate over its feasibility, e.g. Hagen, 2001); but it is more parsimonious to *assume* that it exists than to assume that our pseudonomenology is drastically wrong and requires radical alteration.

Continuing this theme, it is interesting to ask what is the best theory of consciousness we have *now*? Scientists have a duty to provide the *most likely* answers to the public's questions (and at private dinner parties!), even if these answers are by no means certain. In this spirit, I will give a brief overview of an admittedly extremely speculative answer, but which appears to be the best fit we have for our current data (this is essentially the Penrose-Hameroff theory, but with our concept model tacked on):

The brain is a quantum computer, and quantum effects are responsible for the apparent randomness in its behavior which neuroscience is currently unable to account for. Areas of brain involved in awareness constantly enter into superposed states as a means of efficiently performing parallel computations. Microtubules are the machinery for doing this. Superposition allows the brain to entertain and evaluate multiple future situations at the same time. Like an ordinary quantum computer, the superposition becomes reduced to give the result of the calculation. But the type of reduction may be different from that in normal quantum computers, in that the components involved bring about their own reduction, rather than requiring an external source. *This type of reduction is identical to a moment of consciousness, and the information structures contained in the reducing area are the structures of phenomenology.* These reductions occur at around 40-50Hz, and cause, or are caused by, events including NMDA receptor activity and macroscopic γ -oscillations.

The above should be taken with an extremely large grain of salt, resting as it does on a tower of Buddhist reports, untested quantum gravity theory, continental philosophy, immature quantum computation theory, and debatable hypotheses about microtubules' coherence abilities! But it is the best theory we have, and the only reason that the tower is flimsy is for lack of *empirical research*. There are no *philosophical* problems with it³. As with our concept theory, it is a story to be held up for scrutiny, testing and refinement by the scientific community. It may be flimsy now, but it is a starting point. If Buddhist reports, quantum gravity and the rest are taken seriously – not just assumed to be correct, but rigorously challenged, defended and reformulated – then we can build a tower just as strong as that of any other science. We just have to do the work.

Phenomenological simulation *versus* Artificial Consciousness

The concept model we have proposed was designed to be implementable, and was originally intended as a merely *passive* system for subjects to accurately report their phenomenal experiences. We have noted that the content of phenomenal world appears to be passive: we are *not* conscious of the *processes* acting on that content. The content corresponds only to the 'working memory buffer' in Baars' Theatre of Awareness model, for example.

However, it would be intriguing to add an 'unconscious' *active* component to the concept model, to create artificial (non-conscious) awareness. This would not involve starting from scratch – many years of research have been devoted to the kind of cognitive processes acting in awareness. These *processes* could be combined with our model's *data structures* to create a new model of awareness. A system of this kind should be able to perform tasks such as concept learning, analogical reasoning and general problem solving. This would not be a major revolution in cognitive science; rather a way of building on the work of others by augmenting it with our concept structures. The big difference from other models, however, would be that this new model would be *specifically* designed to mirror the contents of human *phenomenology* during tasks. This is something which is conventionally ignored by most current cognitive models.

³ At least to the best of my knowledge. I welcome debate over potential objections that others will undoubtedly raise!

Infinitely more exciting, however, is the prospect of creating true *artificial consciousness*. By this I do not mean ‘artificial awareness’ or ‘artificial high-level cognition’ – but genuine, subjective, phenomenal consciousness of the kind that I *am*. If we are confident that we have found a pseudonomenal identity for consciousness, all we have to do is built an artefact which contains that pseudonomenon. According to our current best theory, this would mean building an artefact with objective quantum state reduction. For the artefact to have a similar kind of phenomenology to us, we could begin with the active-concept system described above, but *implement it as a quantum computer* using the same kind of reduction as the brain.

Practical quantum computer hardware is still about 10 years away (MIT currently has a working 2-qubit device⁴), and the P&H theory suggests that even these machines will lack the particular *kind* of reduction (**Orch-OR**) that is to be identified with consciousness. However, we can imagine that technology will eventually reach the ability to engineer artificial **Orch-OR** – that task is just a *practical* one. However, for now it is already possible to *simulate* quantum algorithms acting on our concept model. Such a simulation would not be conscious, but would allow us to prepare a system ready for the arrival of the conscious hardware.

Such a system might include the quantum AI tree-searcher discussed earlier (p.75). It would be especially interesting to explore how this could be implemented as a superposable neural network, with each neuron or assembly being a concept detector/instantiator as we hypothesised in our model (p.73).

If this sounds far-fetched, we should note that quantum cognition simulations are *already* beginning to appear: Chen (forthcoming) has simulated a quantum natural language processor and nonmonotonic/counterfactual reasoning system which already ‘delivers quite remarkable results’ due to its use of massive parallelisms unavailable to classical systems.

I suggest (and intend to personally pursue) the following *immediate* research strategy for the development of artificial consciousness:

- 1) gain a thorough understanding of quantum computation theory
- 2) implement the passive phenomenology report system; give it to empirical phenomenologists⁵
- 3) combine the system with cognitive theories of awareness processes to create an *active* system
- 4) constantly obtain feedback from phenomenologists and cognitivists, and refine the model
- 5) research how these processes can be efficiently implemented as quantum algorithms
- 6) implement an active simulation of the *active* quantum system
- 7) research **Orch-OR** theory, investigate ways of using it in algorithms
- 8) simulate an active **Orch-OR** version of the system
- 9) wait for the hardware to arrive, then build the artificially conscious machine
- 10) follow changes in pseudonomenology during the above stages, and refine them appropriately.

Difficult answers

This thesis began by asking some ‘difficult questions’, and will now end by answering them. In this chapter, we have drawn three different types of conclusion: *logical*, *empirical* and *speculative*. (The latter being attempts to make inferences to the best explanation from currently sparse data. The boundary between empirical and speculative findings is thus somewhat blurry.) I use these same labels to indicate confidence in the following answers. The questions fall into two groups: the first about consciousness itself, and the second about its SELF:

How does consciousness fit into our physical picture of the universe?

Logically: We maintain a pseudonomenal model of the fundamental nomina of the universe, and we must identify consciousness with something in that model. The identifier can either be a pseudonomenon already in the model, or a new specially-postulated pseudonomenon.

⁴ Joseph Chen (University of Hamburg), private conversation.

⁵ Some of whom have already expressed interest in the system.

Empirically: Our early 21st-century pseudonomenology includes quantum state reduction, and this appears to be a parsimonious identity for consciousness. Alternatively, the pseudonomenology could be reformulated to identify consciousness with something else. Empirical correlate searching is needed to determine which of the above is correct, though the former should be assumed in the meantime.

Speculatively: Consciousness is a particular type of quantum state reduction occurring (so far) only in the brain.

Are other people conscious, and how can we tell?

Logically: We can never know the nomenal world for certain, but our pseudonomenal model postulates the existence of an external reality which gives useful and accurate predictions. This model is our best available knowledge of the true structure of nomenality. If the model declares others to be conscious, then we should accept this fact.

Empirically: Our pseudonomenology does not give *me* any special status over other people – my own brain is supposed to have the same components as those of others. So given that I am conscious, I should accept that others are too.

Which animals are conscious, and how can we tell?

Logically: Again, our pseudonomenology should be trusted to provide the answer to this question.

Speculatively: According to the Penrose-Hameroff model – our best speculative theory – animals are conscious if they display orchestrated reduction in their brains. This theory has calculated that the presence and frequency of conscious moments depends on the mass of the brain, initially appearing in creatures about the size of the *C elegans* worm. Larger brains have more moments per second, and, according to our concept theory, these moments are *bigger*, containing more phenomena. The P&H equations suggest that an alert cat might have roughly the same amount of consciousness as a human lying on a sofa on a lazy Sunday afternoon after eating a big lunch.

How do anaesthetics work?

Logically: By inhibiting the pseudonomenon that is consciousness.

Empirically: We have seen (p.28) that anaesthetics have been shown to prevent quantum superposition in microtubules. According to the P&H model, this inhibits the quantum reduction pseudonomenon.

Can an appropriately programmed Turing machine be conscious, and if so how?

Logically: No. Turing machines and their functions exist only phenomenally; consciousness exists nomenally, so they can never be identical.

Can any man-made artefact become conscious, and if so, how?

Logically: Yes. Baby-making aside, we ‘simply’ need to locate the pseudonomenon that is identical to consciousness, and build an artefact containing it.

Speculatively: This can be done under our current pseudonomenology by executing the immediate research programme suggested earlier.

Could we bring a dead person back to consciousness? Would he be the same person?

Empirically: This question concerns the continuity of the *self*. The question asks: “If we could return a brain to a previous (i.e. non-dead) state, would its consciousness be the same one?” This is in fact a non-question: we have seen that consciousness itself is not continuous, but a series of discrete moments of phenomenology. The only reason that one appears to experience continuity is the presence of the SELF concept, which stores the history of a body’s behavior. But in fact each moment is an isolated event; a different consciousness. ‘I’ am different at each successive moment, by my SELF stays (almost) the same. So ‘I’ – my consciousness – actually ‘die’ and ‘become reborn’ fifty times per second during normal life! A related, replacement question would thus be: “If we could return a brain to a previous (i.e. non-dead) state, would its SELF be the same one, as in normal life?”, and the answer would be *yes*, because a brain in the same state would contain the same SELF concept.

Could we transfer a person from one body to another, or into a machine?

Empirically: This is essentially the same question as above: the answer is yes, if (a) we copy the SELF concept and (b) the new body or machine has the consciousness identity acting on that SELF concept.

Could we build a teleporter ... should I agree to use it?

The device is supposed to work by building a copy of me, then destroying the original. There are two questions here: (1) how do we know that the teleporter will produce a *conscious* copy rather than an unconscious functional replica (a 'zombie')?; and (2) even if the copy is conscious, will it be *me*, or will it be someone else, leaving me dead? We will answer these questions separately:

(1.) *Logically:* As we can never be certain of the correctness of our pseudonomenclature, we can never be certain that a putative teleporter actually works: it *might* just be making a zombie replica. If that was the case, then using it would be a very bad idea. However, although we can never be certain of our pseudonomenclature we can become increasingly *confident* in it. Our confidence in it should be given by the degree of phenomeno-physical correlation that our science has observed. We can never be certain that aeroplanes will fly, but our science has observed so many correlations concerning the laws of flight that we agree to use them. In exactly the same way, observing enough phenomenal-objective correlations will give us a similar level of confidence in the safety of the teleporter. The teleporter would become as safe as aeroplanes.

(2.) *Empirically:* The answer is *yes*, for the same reason as the answers to previous questions. I.e. Consciousness is momentary and discrete, but it is the SELF that is teleported and which we should be concerned about. If the SELF is successfully teleported then yes, it would still be 'me'.

What if the teleporter did not destroy the original? Which copy would be the real me?

Empirically: This is a non-question. No two conscious moments are the same – my experience of continuity is an illusion. In this situation, there would be two new consciousnesses, each of whose experiences contain the same SELF concept. Both would experience the illusion of continuity from the original 'me'. But the original 'me' would no longer exist, just as it no longer exists after *any* moment has passed.

The question is often rephrased: 'if before stepping into the copying device, I have to sign a contract saying that one or other copy should be destroyed, then which copy should I choose?'. Evolutionarily, we have a natural instinct for SELF-preservation – our goal is to maintain the embodied existence of the SELF concept. But in this case, there are now two SELF embodiments! Our pretheoretical inclination is probably to destroy the newly-made copy, but in fact both copies are equally conceivable as 'the future me' so there really is no difference. If we wish to fulfil our evolutionary urge to survive and reproduce, the best choice is to renegotiate the contract and allow both copies to survive!

Can we become immortal, preserving our consciousness indefinitely?

Empirically: Again, this is a non-question – what is preserved from one moment to the next is the SELF, not consciousness. The real question here is 'can we preserve the SELF indefinitely?'. Aside from anti-ageing medicines, there are two conceivable ways to do this. First, transfer the SELF concept from body to body, or to machine, as discussed above. This already happens to a small extent: Jesus and Elvis' SELFs live on the bodies of psychiatric patients who are deluded into thinking that they *are* one of those people: they are right, in our sense! Secondly, we may *reconceive* the SELF as something other than our own particular physical body – as happened above in the non-destructive teleporter example.

But there are less technological ways to reconceive the self. We may for example re-identify it with our *family* – and live on as our children. Strong family identity is common in traditional aristocratic families, and gives a justification for inheritance laws: it is fair that a self should possess a large sum of money if the same self worked hard to earn it hundreds of years ago. Alternatively, we may re-identify our self with our country or tribe or religious group: martyrs are willing to sacrifice their physical body for the continuity of their communal self. A great farmer conceives of his land as part of his self; his mere body can pass away happy in the knowledge that his work on the soil still lives. Similarly, we may identify our self with our ideas, our words and our dreams, to live on forever in the minds of others, passed on from each generation of thinkers to the next.

References

- Anderson, J. (1993). *Rules of the Mind*. Hillsdale, NJ: Lawrence Erlbaum Associates Inc.
- Anscombe, G.E.M. (1960). *The First Person*.
- Baars, B. (1988). *A Cognitive Theory of Consciousness*. Cambridge: CUP.
- Baars, B. (1999). 'There is already a field of systematic phenomenology, it's called psychology'. *Journal of Consciousness Studies* 6:213-215.
- Berkeley, G. (1713). *Three Dialogs between Hylas and Philonous*. Reprinted 1975, in *Philosophical Works*. London: Everyman.
- Blackburn, S. (1984). *Spreading the Word*. Oxford: Clarendon.
- Block, N. (1978). 'Troubles with functionalism'. In *Perception and Cognition*, Vol. IX. Ed. Savage, W. Minnesota: University of Minnesota Press.
- Block N., Flanagan, O. and Güzeldere, G. (1997). *The Nature of Consciousness: Philosophical Debates*. Cambridge, MA: MIT.
- Bohm, D. (1952) 'A suggested interpretation of the quantum theory in terms of hidden variables.' I and II, *Phys. Rev.* 85:166-93.
- Bohr, N. (1934). *Atomic Physics and Human Knowledge*. Cambridge: CUP.
- Borodistsky, L. (2001). 'The roles of body and mind in abstract thought.' *Proc. 23rd Ann. Conf. Cog. Sci. Soc.* (Edinburgh University, 2001).
- Boring, E. (1942). *Sensation and Perception in the History of Experimental Psychology*. New York: Appleton.
- Boring, E. (1953). 'A history of introspection'. *Psychological Bulletin* 50:169-189.
- Braddock, G. (in press). 'Beyond reflection in naturalised phenomenology'. *Journal of Consciousness Studies*, November 2001.
- Bray, D. (1995). Protein molecules as computational elements in living cells'. *Nature* 376:307-312.
- Brentano, F. (1874). *Psychology from an Empirical Standpoint*. Leipzig. English trans. McAlister, L., 1974. London: Routledge.
- Burgess, N. and O'Keefe, J. (1996). 'A model of hippocampal function'. *Neural Networks* 7:1065-1081.
- Cage, J. (1939). *Silence*. Repr. 1995, London: Marion Boyars.
- Campbell, K. (1990). *Abstract Particulars*. Oxford: Blackwell.
- Chalmers, D. (1995). 'Facing up to the problem of consciousness'. *Journal of Consciousness Studies* 2:200-219.
- Chalmers, D. (1996). *The Conscious Mind*. Oxford: OUP.
- Chalmers, D. (1997). 'Moving forward on the problem of consciousness'. *Journal of Consciousness Studies* 4:3-46.
- Chalmers, D. (2000). 'What is a neural correlate of consciousness?'. In Metzinger, 2000, Chapter 2.
- Chen, J. (forthcoming). *Quantum Computation and Natural Language Processing*. PhD thesis, Computer Science Department, University of Hamburg. (A paper is also under submission to *Cognitive Science*.)
- Churchland, P.S. (1996). 'The hornswoggle problem'. *Journal of Consciousness Studies* 3:402-408..

REFERENCES

- Colthart, M., Curtis, B., Atkins, P. and Haller, M. (1993). 'Models of reading aloud: Dual-Route and Parallel-Distributed Processing approaches'. *Psychological Review* 100:589-608.
- Crick, F. (1994) *The Astonishing Hypothesis: The Scientific Search for the Soul*, London: Simon and Schuster.
- Crick, F. and Koch, C. (1990). 'Towards a neurobiological theory of Consciousness', *The Neurosciences* 2:263-275. Crick, F. and Koch, C. (1995) 'Visual processing in V1 is not conscious'. *Nature* 375:121-123.
- Crick, F. and Koch, C. (1998). 'Consciousness and Neuroscience', *Cerebral Cortex* 8:97-107.
- De Valois R. and De Valois, K. (1975). 'Neural coding and color'. In E. Carterett & M. Friedman (eds.), *Handbook of Perception* (Vol. 5). New York.
- Dennett, D.C. (1987). *Brainstorms: Philosophical essays on Mind and Psychology*. Cambridge, MA: MIT.
- Dennett, D.C. (1991). *Consciousness Explained*. Penguin.
- Dennett, D.C. (1991b). 'Real Patterns'. *Journal of Philosophy* 88:27-51.
- Descartes, R. (1641). *Meditations on First Philosophy*. Trans. D. Cress, 1993. Cambridge: Hackett.
- Edelman, G.M. & Tononi, G. (2000a). 'Reentry and the Dynamic Core'. In Metzinger, 2000, Chapter 9.
- Edelman, G.M. & Tononi, G. (2000b). *A Universe of Consciousness*. Penguin.
- Eichembaum, D. (2000). 'A cortical-hippocampal system for declarative memory'. *Nature* 2000(1).
- Evans, J.M. (1987). Patients' experiences of awareness during general anaesthesia. In Rosen & Lunn, p.84-192.
- Everett III, H. (1957). 'Relative state formulation of quantum mechanics'. *Rev. of Mod. Phys.* 29:454-462.
- Eysenck, M.W. and Keane, M.T. (2000). *Cognitive Psychology*. (4th Ed.) Sussex: Psychology Press.
- Flohr, H. (2000). NMDA 'Receptor-Mediated Computational Processes and Phenomenal Consciousness'. In Metzinger, 2000, p245.
- Fodor, J. (1983). *The Modularity of Mind*. Cambridge, MA: MIT.
- Fodor, J. (1987). 'Why there still has to be a Language of Thought'. In *Psychosemantics*. Cambridge, MA: MIT.
- Fox, C. (1999). 'Originetix: an evolving artificial life virtual world'. Company confidential tech. report. Cyberlife Technology, Cambridge.
- Fox, C. (2001). 'Concepts: structures and processes'. MSc Cognitive Science essay, Edinburgh University. Online at www.charlesfox.org.uk.
- Franklin, S. (2001). '"Conscious" software: the quest for the ultimate artefact'. In *Proceedings of the 23rd Conference of the Cognitive Science Society*. (University of Edinburgh).
- Franks, N.P. and Lieb, W.R. (1982). 'Molecular mechanisms of general anaesthesia.' *Nature* 274:339-342.
- Franks, N.P. and Lieb, W.R. (1984). 'Do general anaesthetics act by binding competitively to specific receptors?' *Nature* 310:599-601.
- Franks, N.P. and Lieb, W.R. (1998). 'Which molecular targets are most relevant to general anaesthesia?' *Toxicology Letters* 100-101:1-8.
- Franks, N.P. and Lieb, W.R. (2000). 'What can we learn from Anaesthetic Mechanisms?' In Metzinger, 2000, p265.
- Gallagher, S. (1997). 'Mutual Enlightenment: Recent phenomenology in cognitive science'. *Journal of Consciousness Studies* 4:195-214.
- Gardner, H. (1985). *The Mind's New Science: A History of the Cognitive Revolution*. Basic Books.

REFERENCES

- Gray, C.M., Konig, P., Engel, A.K. and Singer, W. (1989) 'Stimulus-specific neuronal oscillations in orientation columns of cat visual cortex'. *Proc Natl Acad Sci USA* 86:1696-1702.
- Güzeldere, G. (1997). 'Approaching Consciousness'. In Block, Flanagan & Güzeldere 1997.
- Hagan, S. (2001). 'Quantum models of consciousness in microtubules: Decoherence and the issue of biological feasibility'. *Consciousness Research Abstracts. Toward a Science of Consciousness 2001 conference proceedings*. Imprint Academic.
- Hebb, D.O. (1949). *The Organisation of Behaviour*. New York: Wiley.
- Hameroff, S. (1996). 'More neural than thou'. *Towards a Science of Consciousness II: The 1996 Tucson Discussions and Debates*. (1998). Ed. Hameroff, Kasziack, Scott. Cambridge MA: MIT.
- Hameroff, S. (1998a). 'Fundamentality': is the conscious mind subtly linked to a basic level of the universe?' *Trends in Cognitive Sciences* 2:119-127.
- Hameroff, S. (1998b). 'Anaesthesia, consciousness and hydrophobic pockets – a unitary quantum hypothesis an anaesthetic action'. *Toxicology Letters* 100-101:31-39.
- Hameroff, S. (1998c). 'Reply to Spier and Thomas from Stuart Hameroff'. *Trends in Cognitive Sciences* 2:125-126.
- Hameroff, S. (1999). 'The neuron doctrine is an insult to neurons'. *Behavioral and Brain Sciences* 22:838-839.
- Hameroff, S. (2001). 'What is Consciousness – Slide show lecture'. Online at www.consciousness.arizona.edu/Hameroff/
- Hameroff, S. & Penrose, R. (1996). 'Conscious Events as Orchestrated Space-Time Selections'. *Journal of Consciousness Studies* 3:36-53.
- Heisenberg, W. (1953). *Physics and Philosophy*. Chapter 8. New York: Harper and Row.
- Hofstadter, D.R. (1985). *Metamagical Themas*. New York: Basic Books.
- Hofstadter, D.R. (1997). *Le Ton Beau de Marot*. New York: Basic Books.
- Hofstadter, D.R. and Dennett, D.C. (1981). *The Mind's I*. New York: Basic Books.
- Honderich, T. (2001). 'Consciousness as cells, as stuff in the head, and as existence.' *Consciousness Research Abstracts. Toward a Science of Consciousness 2001 conference proceedings*. Imprint Academic.
- Hume, D. (1748). *An Enquiry Concerning Human Understanding*. Repr. 1975, Oxford: Clarendon.
- Husserl, E. (1972). *Phenomenology*. Article for the *Encyclopaedia Britannica*. Eng. Trans. *Journal of the British Phenomenology Society* 2:77-90.
- Hut, P. (1999). 'Theory and Experiment in Philosophy'. *Journal of Consciousness Studies* 6:241-244.
- Hut, P. and Shepard, R.N. (1997). 'Turning the hard problem upside down and sideways.' In *Explaining Consciousness: The Hard Problem*. Ed. Shear, J. Cambridge MA: MIT.
- James, W. (1890). *The Principles of Psychology*. Reprinted 1981, Cambridge MA: Harvard.
- Jevtovic-Todorovic, V., Todorovic, S.M., Mennerick, S., Powell, S., Dikranian, K., Benshoff, N., Zorumski, C.F. and Olney, J.W. (1998). 'Nitrous oxide (laughing gas) is an NMDA antagonist, neuroprotectant and neurotoxin'. *Nature Medicine* 4:460-463.
- Levine, J. (1983). 'Materialism and qualia: the explanatory gap'. *Pacific Philosophical Quarterly* 64:354-61.
- Leibniz, G. (1695). *New System of Nature and the Communication of Substances*. Trans. G. Parkinson & M. Moris, 1973. London: Dent.
- Llinas, R.R. and Pare, D. (1991). 'Of dreaming and wakefulness'. *Neuroscience* 44:521-535.
- Locke, J. (1690). *An Essay Concerning Human Understanding*. Reprinted 1975. Oxford: Clarendon.
- Lumer, E.D. and Rees, G. (1999). 'Covariation of activity in visual and prefrontal cortex associated with subjective visual perception'. *Proc. Natl. Acad. Sci. USA* 96:1669-1673.

REFERENCES

- Marr, D. (1982). *Vision*. Cambridge, MA: MIT.
- Kant, I. (1781). *Critique of Pure Reason*. Trans. N. K. Smith, 1929. London: McMillan.
- Kiel, F. and Batterman, N. (1984). 'A characteristic-to-defining shift in the development of word meaning'. *Journal of verbal learning and verbal behavior* 23:221-236.
- Kelly, S.D. (in press). 'The non-conceptual content of perceptual experience: Situation dependence and fineness of grain'. Forthcoming in *Philosophy and Phenomenological Research*, 2001.
- Klawiter, A. (2001). 'Auditory consciousness of objects'. *Consciousness Research Abstracts. Toward a Science of Consciousness 2001* conference proceedings. Imprint Academic.
- Kohonen, T. (1982). 'Self-organising formation of topologically correct feature maps.' *Biological Cybernetics* 43:59-69.
- Kosslyn, S., Ball, T. and Reiser, B. (1978). 'Visual images preserve metric spatial information'. *Journal of Experimental Psychology: Human Perception and Performance* 4:47-60.
- Kuhn, T. (1962). *The Structure of Scientific Revolutions*. Chicago: University of Chicago Press.
- MacLennan, B. (1996). 'The elements of consciousness and their neurodynamical correlates'. *Journal of Consciousness Studies* 3:409-24.
- Mangan, B. (1993). 'Taking phenomenology seriously: The "Fringe" and its implications for cognitive research'. *Consciousness and Cognition* 2:89-107.
- Marbach, E. (1993). *Mental Representations and Consciousness: Towards a Phenomenological Theory of Representation*. Dordrecht: Kluwer Academic.
- McClelland, J.L., McNaughton, B.L., O'Reilly, R.C. (1994). 'Complementary Learning Systems in the hippocampus and neocortex: insights from the successes and failures of connectionist models of learning and memory'. Tech. Rep. PDP.CNS.94.1, Carnegie Mellon University.
- McFarlane, C., Warner, D.D., Todd, M.M. and Nordholm, L. (1992). 'AMPA receptor competitive antagonist reduces halothane in rats'. *Anesthesiology* 77:1165-1170.
- McGinn, C. (1989). 'Can we solve the mind-body problem?' *Mind* 98:891, 349-366.
- Metzinger, T. (ed) (2000). *Neural Correlates of Consciousness*. Cambridge MA: MIT.
- Miller, G.A. (1956). 'The magical number seven, plus or minus two: Some limits on our capacity for processing information'. *Psychological Review* 63:81-97
- Nagel, T. (1974). 'What is it like to be a bat?'. *Philosophical Review* 83:435-51.
- Nagel, T. (1986). *The View from Nowhere*. Oxford: OUP.
- Naudin, J., Cros-Azorin, C., Mishara, A., Wiggins, O.P., Schwartz, M.A., Azorin, J. (1999). 'The use of Husserlian reduction as a method in psychiatry'. *Journal of Consciousness Studies* 6:155-171.
- Neilsen, M.A. and Chuang, I.L. (2000). *Quantum Computing and Quantum Information*. Cambridge: CUP.
- O'Regan, K. (2001). 'Experience is not something we feel but something we do'. Online at <http://nivea.psych.univ-paris5.fr/ASSChtml/Pacherie4.html>.
- O'Reilly, R.C. & Rudy, J.W. (2000). 'Computational principles of learning in the neocortex and hippocampus'. *Hippocampus* 10:389-397.
- Paul, D. U. (1981) 'The structure of consciousness in Paramārtha's purported trilogy'. *Philosophy East and West* 31:297-319.
- Penrose, R. (1989). *The Emperor's New Mind*. Oxford: OUP.
- Penrose, R. (1994). *Shadows of Mind*. Oxford: OUP.
- Penrose, R. and Hameroff, S. (1995). 'What Gaps? Reply to Grush and Churchland'. *Journal of Consciousness Studies* 2:99-112.
- Plato. (c.380 BC). *Phaedo*. Reprinted in *The Dialogs of Plato*, 1892. Oxford: Clarendon.
- Popper, K. (1935). *The Logic of Scientific Discovery*. London: Hutchinson.

REFERENCES

- Purves, D., Augustine, G.J., Fitzpatrick, D., Katz, L.C., LaMantia, A.S., McNamara, J.O. (1997) *Neuroscience*. Sunderland, MA: Sinauer.
- Quine, W.V.O. (1953). 'Two dogmas of empiricism'. *From a Logical Point of View*. Cambridge MA: Harvard.
- Rafal, R.D. and Posner, M.I. (1987). 'Deficits in human visual spatial attention following thalamic lesions'. *Proc Natl Acad Sci USA* 84:7349-7353
- Ramscar, M. & Hahn, U. (1998). 'What family resemblances are not: The continuing relevance of Wittgenstein to the study of concepts and categories'. *Proceedings of the 20th Annual Conference of the Cognitive Science Society*, pp865-70. University of Wisconsin: Madison.
- Rang, H.P., Dale, M.M., and Ritter, J.M. (1995). *Pharmacology*. 3rd Ed. Edinburgh: Churchill Livingstone.
- Revonsuo, A. (1995). 'Consciousness, dreams and virtual realities'. *Philosophical Psychology* 8:35-57.
- Revonsuo, A. (2000). 'Prospects for a science of consciousness'. In Metzinger, 2000.
- Revonsuo, A. (2001). 'Can fMRI discover consciousness in the brain?' *Journal of Consciousness Studies* 8:3-21.
- Richter, W., Richter, M., Warren, W.S., Merkle, H., Andersen, P., Adriany, G. and Ugurbil, K. (2000). 'Functional Magnetic Resonance Imaging with Intermolecular Multiple-Quantum Coherences'. *Mag. Res. Imaging* 18:489-494.
- Rosen, M. and Lunn, J.N. (1987). *Consciousness, awareness and pain in general anaesthesia*. London: Butterworth.
- Ryle, G. (1950). *The Ghost in the Machine*. Oxford: OUP.
- Sachs, O. (1985). *The Man who Mistook his Wife for a Hat*. London: Duckworth.
- Schank, R. C. (1972). 'Conceptual dependency: a theory of natural language understanding'. *Cognitive Psychology* 3:552-631.
- Schweitzer, P. (1993) 'Mind/Consciousness Dualism in Sāṅkya-Yoga Philosophy'. *Philosophy and Phenomenological Research* 53:845-859.
- Schweitzer, P. (1996). 'Physicalism, functionalism and conscious thought'. *Minds and Machines* 6:61-87.
- Schopenhauer, A. (1819). *The World as Will and Idea*. Trans. R. Haldane and J. Kemp, 1883. London: Routledge.
- Schwender, D.S., Madler, C., Klasing, S., Peter, K. and Poppel, E. (1994). 'Anaesthetic Control of 40Hz brain activity and implicit memory'. *Consciousness and Cognition* 3:129.
- Searle, J. (1980). 'Minds, brains and programs'. *Behavioral and Brain Sciences* 1:417-424.
- Searle, J. (1992). *The Rediscovery of the Mind*. Cambridge, MA: MIT.
- Seidenberg, M.S. and McClelland, J.L. (1989). 'A distributed, developmental model of word recognition and meaning'. *Psychological Review* 96:523-568.
- Shear, J. (1996). 'Closing the empirical gap'. *Journal of Consciousness Studies* 3:359-375.
- Shear, J. & Jevning, R. (1999). 'Scientific exploration of meditation techniques'. *Journal of Consciousness Studies* 6:189-209.
- Singer, P. (1979). *Practical Ethics*. Cambridge: CUP.
- Sokolowski, R. (2000). *Introduction to Phenomenology*. Cambridge: CUP.
- Spier, E. and Thamas, A. (1998). 'Response from Emmet Spier and Adrian Thomas'. *Trends in Cognitive Sciences* 2:124-125.
- Spinoza, B. (1677). *The Geometry of Ethics*. Netherlands: Den Haag.
- Stapp, H.P. (1996). 'The Hard Problem: A Quantum Approach'. *Journal of Consciousness Studies* 3:194-210.

REFERENCES

- Strawson, P. (1962). 'Freedom and resentment'. *Proceedings of the British Academy* 48:1-25.
- Titchener, E.B. (1898). *Outline of Psychology*. New York: MacMillan.
- Tomlin, S.L., Jenkins, A., Lieb, W.R. and Franks, N.P. (1998). 'Stereoselective effects of etomidate optical isomers on gamma-aminobutyric acid type A receptors and animals'. *Anaesthesiology* 88:708-817.
- Turing, A.M. (1950). 'Computing Machinery and Intelligence'. *Mind* 59:433-460.
- Vaughan, T.M. (ed.) (2000). *IEEE Transactions on rehabilitation engineering*. Special issue on Brain-Computer Interfacing. Vol. 8.
- Velmans, M. (2000). *Understanding Consciousness*. London: Routledge.
- Von der Malsburg, C. and Schneider, W. (1986). 'A neural cocktail party processor'. *Biological Cybernetics* 54:29-40.
- Wallace, B.A. (1999). 'The Buddhist tradition of Samanatha'. *Journal of Consciousness Studies* 6:175-187.
- Wallace, B.A. (2001). 'Intersubjectivity in Indo-Tibetan Buddhism'. *Journal of Consciousness Studies* 8:209-30.
- Walleczek, J. (1995). 'Magnokinetic effects on radial pairs: a possible paradigm for understanding sub kT magnetic field interactions with biological systems'. In *Electromagnetic Fields: Biological Interactions and Mechanisms* (Blank, M., ed.). American Chem. Soc. Books.
- Welch, R.B. (1978). *Perceptual Modification: Adapting to Altered Sensory Environments*. New York: Academic Press.
- Wheeler, J.A. (1990) 'Information, physics, quantum: the search for links'. In *Complexity, Entropy and the Physics of Information*, ed. W. Zurek. Redwood City, CA: Addison Wesley.
- Whitehead, A. (1933). *Process and Reality*. New York: Macmillan.
- Wigner, E. (1961). 'Remarks on the mind-body problem'. In *The Scientist Speculates* ed. I.J. Good. London: Heineman.
- Wittgenstein, L. (1953). *Philosophical Investigations*. Trans. G.E.M. Anscombe, 1968, New York: Macmillan.
- Young, S. (2000). 'Applications of mindfulness meditation in the study of human consciousness'. Abstract from *Towards a Science of Consciousness 2000* pre-conference workshop.