INFORMATION TO USERS

This manuscript has been reproduced from the microfilm master. UMI films the text directly from the original or copy submitted. Thus, some thesis and dissertation copies are in typewriter face, while others may be from any type of computer printer.

The quality of this reproduction is dependent upon the quality of the copy submitted. Broken or indistinct print, colored or poor quality illustrations and photographs, print bleedthrough, substandard margins, and improper alignment can adversely affect reproduction.

In the unlikely event that the author did not send UMI a complete manuscript and there are missing pages, these will be noted. Also, if unauthorized copyright material had to be removed, a note will indicate the deletion.

Oversize materials (e.g., maps, drawings, charts) are reproduced by sectioning the original, beginning at the upper left-hand corner and continuing from left to right in equal sections with small overlaps.

Photographs included in the original manuscript have been reproduced xerographically in this copy. Higher quality 6" x 9" black and white photographic prints are available for any photographs or illustrations appearing in this copy for an additional charge. Contact UMI directly to order.

Bell & Howell Information and Learning
300 North Zeeb Road, Ann Arbor, MI 48106-1346 USA

UMI®
800-521-0600

UNIVERSITY OF CALIFORNIA
Santa Barbara

"Russell, Hayek, and the Mind-Body Problem"

A Dissertation submitted in partial satisfaction
of the requirements for the degree of

Doctor of Philosophy

in

Philosophy

by

Edward Charles Feser

Committee in charge:

Professor C. Anthony Anderson, Chairman

Professor Anthony Brueckner

Professor Matthew Hanser

June 1999

The dissertation of Edward Charles Feser is approved

Anthony Bruckner

Matthew Hunt

C. Anthony Anderson

Committee Chairman

June 1999

June 20, 1999

ACKNOWLEDGEMENTS

For valuable comments on drafts of this dissertation and/or on ancestors of some of its chapters, I thank C. Anthony Anderson, Anthony Brueckner, David Chalmers, Francis Dauer, Kevin Falvey, Michael Lockwood, Hubert Schwyzer, and Galen Strawson.

For her patience and encouragement over the many years of work that led up to this dissertation, and for so much else, I thank my beloved wife Rachel Feser.

My mother and father, Linda Feser and Edward A. Feser, made my education, and most of what is good in my life, possible. It is with great love and appreciation that I dedicate this work to them.

VITA

April 16, 1968	Born – Los Angeles, CA
1989	A.A., Theology, Ambassador College
1992	B.A., Religious Studies/Philosophy, California State University at Fullerton
1995	M.A., Religion, The Claremont Graduate School
1996	M.A., C.Phil., Philosophy, University of California at Santa Barbara
1996-99	Teaching Assistant, Department of Philosophy, University of California at Santa Barbara. Courses assisted: Introduction to Philosophy, History of Modern Philosophy
1997-99	Lecturer, Department of Philosophy, Pasadena City College. Courses taught: Philosophy of Religion, Contemporary Moral Problems
1998	Teaching Associate, Department of Philosophy, University of California at Santa Barbara. Course taught: Introduction to Philosophy
1998-99	Lecturer, Department of Philosophy, Loyola Marymount University. Course taught: Philosophy of Human Nature

TEACHING AREAS

Areas of Specialization:	Areas of Competence:
Philosophy of Mind	History of Modern Philosophy
Philosophy of Religion	Contemporary Moral Problems
Political Philosophy	

PUBLICATIONS

Without Reason: The Attack on 'Evidentialism' in Contemporary Philosophy of Religion. Unpublished M.A. thesis, Claremont Graduate School, 1994. 95 pp.

"Has Trinitarianism Been Shown to be Coherent?" *Faith and Philosophy*, 14 (1),1997, 87-97.

"Swinburne's Tritheism" *International Journal for Philosophy of Religion*, 42 (3), 1997, 175-184.

"Hayek on Social Justice: Reply to Lukes and Johnston" *Critical Review*, 11 (4), 1997, 581-606.

"Can Phenomenal Qualities Exist Unperceived?" *Journal of Consciousness Studies*, 5 (4), 1998, 405-414.

"Hayek, Social Justice, and the Market: Reply to Johnston" *Critical Review*, 12 (3), 1998, 269-281.

PROFESSIONAL ACTIVITY

"Hayek's Solution to the Mind-Body Problem" Presented at the conference "Toward a Science of Consciousness" at the University of Arizona, Tucson, April 1998. (Abstract published in *Consciousness Research Abstracts: Toward a Science of Consciousness 1998*)

HONORS

Don Gard Award for Academic Achievement in Religious Studies, 1991, California State University at Fullerton

William Alamshah Memorial Prize for Outstanding Essay in Philosophy, 1991-92, California State University at Fullerton

Humane Studies Fellowship 1997-98, sponsored by the Institute for Humane Studies at George Mason University

Doctoral Candidacy Fellowship 1998-99, awarded by the Graduate Division of the University of California at Santa Barbara

ABSTRACT

"Russell, Hayek, and the Mind-Body Problem"

by

Edward Charles Feser

Consciousness, intentionality, and rationality are all features of the mind that philosophers have thought it difficult to account for in naturalistic terms, but it is consciousness that is often considered the most problematic. In particular, how precisely to explain the relationship of qualia, the subjective, first-person features of conscious experience, to the brain (and to the objective, third-person material world in general) is regarded as the central part of the mind-body problem. I argue that materialism and dualism in all their forms have failed to explain this relationship, and that their failure indicates a need to rethink the conceptions of mind and matter typically presupposed by both common sense and philosophical reflection. As Bertrand Russell suggested in some neglected writings, our knowledge of the material world external to the mind is indirect, mediated by our direct awareness of qualia themselves; and what we know of that external world is really only its causal structure rather than its intrinsic nature. The common assumption that matter as it is in itself is utterly unlike mind as revealed in introspection is thus unfounded; in fact, in our introspection of qualia we are directly aware of features of the brain. Dualism thus errors in assuming the mind to exist over and above the brain, but materialism also errors, in assuming that physics and neurophysiology give us a surer grasp of the nature of the brain than does introspection. Despite its insights, the Russellian view also errors, though, in supposing that in our awareness of qualia, at least, we have a grasp of some of the intrinsic qualities of the material world. Following some leads suggested in the work of F.A. Hayek, I argue that even what we know of the internal world of the mind/brain, the sensory order of qualia, is only its structure. In the light of the facts about the nature of our knowledge of the natures of mind and matter, the qualia problem dissolves. Ironically, however, the Hayekian position I defend also implies that the other, on the surface less problematic, features of mind, namely intentionality and rationality, are ultimately inscrutable.

TABLE OF CONTENTS

0. Introduction

(a) Three mind-body problems

The expression: "the mind-body problem," though presumed to name a philosophical perennial and appearing in hundreds of book and article titles, is really a bit of a misnomer, at least if we take seriously the "*the*." For one thing, what counts as a statement of the mind-body problem seems often to depend on the solution one gives to it: It is often introduced to undergraduates as the problem of explaining how mind and body can possibly interact, a formulation which, given the real distinction it implies, makes it sound as if dualism (the view that mind and matter are fundamentally different sorts of thing) were uncontroversial and the issue facing us was merely how to fill in the details; while the way it is often discussed in the current literature implies just the opposite, presupposing that dualism *must* be false, or is at least to be avoided at all costs, the issue being that of determining which way of avoiding it is least implausible. For another, it is clear that there isn't a *single* mind-related phenomenon that is philosophically problematic, but several, as is indicated by the variety of (alleged) paraphrases "the mind-body problem" is given in college course syllabi, middle-brow television documentaries, and the plethora of popular and semi-popular books that have appeared on the subject in the last few years: "Can science explain consciousness?", "Do we have souls?", "Can computers think?", and so forth. So it won't do in introducing the essay that follows merely to note that it concerns the mind-body problem; it must also be made explicit *exactly* what problems we will be

discussing, and more importantly, which problems we will (and which we will not) be hoping to solve.

Jerry Fodor's pithy summation of the tangle of issues that makes up the mind-body problem, as it is understood in recent philosophy, is as good a starting point as any:

> Lots of mental states are *conscious*, lots of mental states are *intentional*, and lots of mental processes are *rational*, and the question does rather suggest itself how anything that is material could be any of these (1994, p. 292).

As Fodor implies, human beings are material systems – or at any rate, they appear to be, both to cursory inspection and to sophisticated scientific inquiry (especially physiological and neuroscientific inquiry). And yet they are associated with certain phenomena for which an explanation in the standard materialistic terms of modern science seems extremely difficult, and which, accordingly, pose a number of philosophical problems:

The problem of *consciousness* is the problem of explaining how a purely material system such as the brain can give rise to the experiences we are all familiar with, how something governed by exactly the same sorts of physical laws that govern obviously non-conscious entities could produce such phenomena as pains, tickles, itches, and the whole range of visual, auditory, tactile, gustatory and olfactory sensations, which are unique to (and definitive of) conscious beings: After all, rocks, tables, and chairs aren't conscious, nor do more complicated material systems like

Buicks, calculators, and even biological systems like digestive tracts and circulatory systems seem to be; so why should the brain, which seems to differ in complexity from these only by degree, produce conscious experiences?

The problem of *intentionality* is the problem of explaining how the property of intentionality, namely the property of "aboutness," of being meaningful or of representing the world as being a certain way, can be had by mental states like beliefs, desires, and the like, especially if the latter are material: After all, natural processes like a river's erosion of a canyon wall, or a tree's production of sap, or a liver's secretion of bile, don't have this property; none of these things "means" or "represents" anything outside itself. So how can we explain why another process of the same general type, namely a brain process, can have this property, e.g. how it can represent the state of affairs of it's being sunny outside – as it would have to do if we are to say that my belief that it's sunny outside is identical with some process taking place in my brain?

The problem of *rationality* is the problem of explaining how, if we are purely material systems, we can be such that we are capable of reliably moving from one thought to another in a manner that corresponds with the laws of logic: After all, other material systems don't seem to be, even though they do uniformly act in accordance with *causal* laws – for instance, planets regularly orbit stars (roughly) in accordance with Kepler's laws, but this doesn't amount to their *reasoning* from, say, "Socrates is a man" and "All men are mortal" to "Socrates is mortal" in a regular way. So why

should brain processes, in bringing about other brain processes in accordance with causal laws, be expected to mirror the laws of logic in any more reliable a manner – even if the *thoughts* that Socrates is a man, and that all men are mortal, and that Socrates is mortal, are identical with processes in the brain?

There are, then, at least *three* mind-body problems, and they all concern the issue of whether and how the mental phenomena exhibited by human beings (and, in some cases, other organisms) can be explained in the same physical terms other features they exhibit can be – which leaves open the possibility that they *can't* be so explained, in which case it might turn out that there is more to human beings (and other organisms?) than the sorts of physical properties they share in common with the other parts of the natural world studied by the various sciences. So any purported solution to "the mind-body problem" must make it clear to *which* of these problems it applies.

The essay that follows will focus on the problem of consciousness, and to that problem alone will it claim to offer a solution. That solution will, however, have implications for the problems of intentionality and rationality, so we will have reason to come to some (tentative) conclusions with respect to them as well, conclusions which are, if incomplete, hopefully also at least suggestive. The solution defended will not be one which could in any usual sense be described as "materialistic" or "physicalistic"; though neither could it be described as a form of dualism or idealism, and it is, in its own way, just as "naturalistic" as physicalism or materialism is. But

then, these terms are fairly slippery in the first place, and we would do well before we proceed to say a little about how they will be used in this essay.

(b) Materialism, physicalism, naturalism

"Materialism" is probably the most familiar of these expressions, and the basic idea it is generally intended to convey can be summed up as follows: Consider the things common sense takes as paradigm cases of physical objects and processes: rocks, tables, chairs, mountains, planets, galaxies, and moving, colliding, heating, melting, freezing, and the like; and consider also the things natural science tells us underlies all these phenomena: molecules, atoms, gravitational force, electromagnetism, and so forth. *Everything* that exists, materialism says, is *like those things* (in their fundamental constitution, say, or in their governing principles); and anything that *seems* not to be like those things either really is like them after all, and can be seen to be so upon adequate investigation, or else doesn't really exist at all.

"Physicalism" is often used as a synonym of "materialism," though sometimes it is used instead to convey the more narrow meaning of materialism about mental phenomena in particular. Moreover, since the account of materialism just given is pretty vague (though it expresses, I think, the working idea behind most materialist philosophy), "physicalism" also seems often to name a tightened up version of the materialist position, and is perhaps more commonly used among philosophers these days than "materialism" is. The idea here is more or less that what should counted as real are just those entities, properties, and processes that physics – or, more plausibly,

a "completed" physics – takes to be the fundamental entities, properties, and processes that underlie everything, or that are at least *reducible* to such entities, properties, and processes.

Of course, even this conception is somewhat vague: For instance, what is meant exactly by "reducible"? Some physicalists would insist on an *ontological* reduction, on which only that which can be shown to *be* "nothing but" the sorts of things physics takes to be fundamental is to be counted as real. Others would insist only on an *explanatory* reduction, on which anything that is to be counted as real must at least be shown to *"supervene"* on fundamental physical properties, where one thing supervenes on another just in case there could not be a difference in the first without a difference in the second: on this view, only what supervenes on fundamental physical facts will turn out to be real, though what supervenes on these facts need not be *identical* to or *ontologically* reducible to such facts.

Probably most physicalists today would opt for the second sense of "reducible," and almost all would take the ontology, not of current physics, but only of a "completed" physics, as the touchstone of reality; but even so, physicalism is still not an entirely determinate position. For at the end of the day everything really hinges on what the ontology of a "completed" physics will include, and it isn't at all clear that this rules much out. As Noam Chomsky has said, "as soon as we come to understand anything, we call it 'physical'" (quoted in Searle, 1992, p. 25), and there have been instances in the history of science where what previous generations would have taken

to be non-physical, and indeed disreputably "occult," entities, properties, and processes came to be regarded as fundamental features of the physical world and essential to physical theory, e.g. the "spooky action at a distance" (to use Einstein's words) which seems to be entailed by quantum mechanics.

Daniel Dennett gives what I think is in fact the working assumption among physicalists of what the future of physics must hold, however, when he writes that any scientific theory of the mind "will have to be constructed from the third-person point of view, since *all* science is constructed from that perspective" and that only third-person data comprises "the data that scientific method permits" (1992, p. 71). The implication here is that, whatever surprising changes in our conceptual scheme the future of physics has in store for us, it *won't* involve the abandonment of an exclusively third-person account of the world. Now the third-person point of view is, roughly, the point of view that can be taken by all inquirers, and third-person data is just that data which is equally accessible to all inquirers – as contrasted with the first-person point of view, which can be taken only by an individual introspecting the contents of his own mind, first-person data being whatever data he introspects there. Much more will be said about this distinction in chapter 1, but suffice it for now to note that the only apparent examples of "first-person" phenomena are *mental* phenomena, so that to say, with physicalism, that only that which is countenanced by a completed physics is real, and that a scientific account of the mind must be constructed

from the third-person point of view, implies that first-person phenomena either can be reduced to third-person phenomena, or don't really exist after all.[1]

That this is what physicalism is committed to is evident from a consideration of the major physicalist theories of mind that have been developed in this century:

Behaviorism (Ryle, 1949) (sometimes called "logical behaviorism" or "philosophical behaviorism" to distinguish it from the behaviorist approach in psychology) held that mental phenomena are really nothing but certain kinds of behavioral phenomena – or, more accurately, that mentalistic *language* was analyzable in terms of language about behavior or behavioral dispositions. So to be in pain, for example, is, on this view, just to exhibit, or be disposed to exhibit under certain conditions, such behavior as wincing, crying out, nursing of the injured part of the body, etc. Behavior and dispositions to behavior are in principle accessible from the third-person point of view; so we see that on the behaviorist view, mental phenomena are, though apparently first-person phenomena, really reducible to third-person phenomena.

The *mind-brain identity theory* (Smart, 1959) is no less obviously committed to a third-person account of the mind: all mental phenomena, on this view, are

---

[1] Compare Fodor (1987, p. 97): "I suppose that sooner or later the physicists will complete the catalogue they've been compiling of the ultimate and irreducible properties of things. When they do, the likes of *spin*, *charm*, and *charge* will perhaps appear on their list. But *aboutness* surely won't; intentionality simply doesn't go that deep. It's hard to see, in face of this consideration, how one can be a Realist about intentionality without also being, to some extent or other, a Reductionist. If the semantic and the intentional are real properties of things, it must be in virtue of their identity with (or maybe of their supervenience on?) properties that are themselves *neither* intentional *nor* semantic. If aboutness is real, it must be really something else."

numerically identical with states or processes in the brain and/or central nervous system, just as water is identical to H2O or lightning is identical to electrical discharges; and this identity can be established in the same way that the latter identities were established by physical science. States and processes in the brain and central nervous system are as accessible from the third-person point of view as H2O, electrical discharges, and the like are; so again, mental phenomena turn out to be a kind of third-person phenomena, despite their appearing to be (at least partially) irreducibly first-person phenomena.

Eliminative materialism (Churchland, 1981) is even more clearly committed to an exclusively third-person picture of the world: mental phenomena, and especially irreducibly first-person mental phenomena, simply do not exist, on this view. They are fictions on a par with the entities postulated by defunct scientific theories of the past, part of a primitive theory, "folk psychology," which may have served mankind well in explaining and predicting human behavior for much of human history (just as Ptolemaic astronomy allowed us to explain and predict the motions of heavenly bodies) but which is nevertheless destined to be superseded or eliminated in favor of a more sophisticated theory which explains behavior in completely neuroscientific and/or cognitive scientific terms (just as Ptolemaic astronomy was superseded by the superior Copernican astronomy). What really cause human behavior are just such third-person phenomena as brain states and processes. Beliefs, desires, and other first-person

phenomena need not be reduced to such third-person phenomena, as the identity theory would have it; for they do not *really* exist *at all.*

Functionalism (Putnam, 1991), probably the most widely accepted theory of mind today, is in fact not intrinsically physicalist; indeed, the position I intend to defend in this essay might be thought of as a *non-physicalist* variety of functionalism. Nevertheless, it is as popular as it is precisely because it is thought to provide a way of characterizing the mental on which an identification of mental phenomena with physical phenomena is unproblematic. The basic idea is that the essential or defining feature of any type of mental state is its functional role, that is, the set of causal relations it bears to (1) environmental effects on the body, (2) other types of mental states, and (3) bodily behavior. So pain, for example, is just that mental state which (1) typically results from bodily damage or trauma, (2) causes distress, annoyance, and practical reasoning aimed at relief, and (3) causes wincing, crying out, and nursing of the injured area. *Anything* playing exactly this causal role just is an instance of pain. This makes the view attractive to physicalists because it seems to allow that mental phenomena can be identified with physical phenomena, since purely physical states and processes, which are third-person phenomena, can play the functional roles in question. Again, the claim is that apparently irreducibly first-person phenomena can be shown to be third-person phenomena.

In what follows, I will thus take physicalism (and materialism, which I will treat as more or less the same doctrine) to be the view that what are (or will turn out to be)

real are just the fundamental features of the physical world which will be uncovered by a completed physics, or features which supervene on these fundamental features, where this is thought to exclude any irreducibly first-person phenomena.[2]

"Naturalism" is another term often used these days, and sometimes it too is used as a synonym for "materialism." But it is also used in less-confined ways, and sometimes even to describe positions which are decidedly non-materialist or non-physicalist. The intuitive idea behind naturalism is that human beings and other creatures with minds are just as much parts of the natural world as lower animals, plants, rivers, rocks, planets, and galaxies are, and that they are as capable of being understood in a broadly scientific way as those other phenomena are.[3] That is, there are no *supernatural* aspects to human beings – souls or immaterial substances, say, or *elan vital* – which are in principle impervious to understanding through scientific inquiry. In this regard, naturalism would entail the denial of the sort of *dualism* associated with Plato and Descartes, on which the mind and self are distinct from the body and brain. Of course, by itself, this sounds pretty close to the descriptions given of materialism and physicalism; but naturalism is understood by many of those who

---

[2] It should be noted, however, that the terms "materialism" and "physicalism," though I do believe they are generally used to denote the sort of position I've characterized, are occasionally used in a very different way. For instance, the Russellian position I'll be discussing in chapters 4 and 5, though it is radically non-materialistic or non-physicalistic in the sense in which I (and I think most writers) use these terms, has been described by some of its proponents as a variety of materialism or physicalism. We'll see later why this is so, and why "materialism" and "physicalism" are nevertheless highly misleading labels for this position.

[3] I should note, however, that there are other approaches to the study of human nature, often called "hermeneutical," which are also often described as types of naturalism, but which reject the scientific tendency to proceed by way of constructing *theories*. Wittgenstein's philosophy, on which an

adopt the label not necessarily to be committed to a denial of the reality of irreducible first-person features. John Searle and David Chalmers, for example, are both philosophers who reject physicalism and its insistence on an exclusively third-person approach, yet who would nevertheless call themselves naturalists. On their view (for reasons we will be looking at in detail later), science and philosophy need not and should not confine themselves dogmatically to a third-person approach.

In this broad sense of "naturalism," the view I will be defending can be considered naturalistic, though again, not materialistic or physicalistic; for while it allows that the mind is as much a part of nature as any other phenomenon studied by the various sciences, and is susceptible of scientific understanding, it also rejects the physicalist's insistence on an exclusively third-person approach. It should be understood, though, that the naturalism I defend here applies *only* to the mind. For naturalism more broadly conceived – like materialism and physicalism – also commits one to rejecting Platonism (the view that there are abstract objects, universals and numbers, say, which exist apart from the physical world) and theism (the view that there is a God), and I take no position on these issues in this essay.

(c) Outline of the argument

The standard approach to the mind-body problem is to start with from the third-person perspective of science and common sense and try to determine how exactly the apparently irreducibly first-person realm of the mind relates to it. I will

---

understanding of the mind and of human nature and human practices in general involves clarification

argue that in fact, all we have any direct knowledge of is the first-person realm of the mind itself, and that all we know even of the states we introspect within it is relatively little, namely the relations these states bear to one another; everything else, including everything we know about the rest of the internal and external worlds, we know only by inference, and even that is knowledge only of structure, not of the world's intrinsic nature. What we do know, however, gives us grounds to identify the mental realm with a portion of the physical realm, namely the brain; and the nature of our knowledge of the world also entails that the barriers usually thought to stand in the way of such an identification are entirely illusory. Illusory too are the physicalist's picture of matter and the dualist's picture of mind.

That's pretty vague as a summary, of course; but so radically different from the standard options is the view that I wish to defend, that a clearer picture can emerge only after the working assumptions of those options are exposed so that the possibility of alternative assumptions, leading to an alternative approach, can be seen. There is no shortcut to the position I will argue for, and the hazy view of our destination I've provided can get clearer only as we proceed on our journey to it. Hopefully the reader is intrigued enough to want to set out on that journey; but if he's still hesitant, the following roadmap might make things easier:

In chapters 1-3 of this essay, I assess the current state of play in the philosophy of mind with respect to the problem of consciousness, particularly what has come to

---

of the way language functions, is one such approach.

be known as the "hard problem" of consciousness, the problem of qualia. Chapter 1 spells out exactly what the problem is. Chapter 2 considers attempts to dissolve the problem, to show it is really a pseudo-problem, and finds them wanting. In chapter 3, I argue that the most widely discussed solutions to the problem of consciousness, namely the various physicalist theories on the one hand, and property dualism on the other, all fail; and that, since these two approaches are widely thought to be the only plausible alternatives, their failure indicates that there are some unexamined – and false – hidden assumptions made by both sides, the rooting out of which will open the way to a genuine solution to the problem of consciousness.

Chapters 4 and 5 deal with one of these false hidden assumptions, the supposition that we have a transparent grasp of the intrinsic nature of the material world, and examine an unusual approach to the mind-body problem, associated with Bertrand Russell, which takes the rejection of this supposition to hold the key to that problem's solution. Chapter 4 defends Russell's view that we have no direct knowledge of the material world external to our minds, and that what we do know of it is only its causal structure, and then spells out the way in which Russell and others take this view of our knowledge of the material world to open the way to a unique, non-physicalist version of the mind-brain identity theory, a version immune to the objections fatal to standard physicalist identity theories. Chapter 5 argues that the Russellian mind-brain identity theory, though an important breakthrough, has insuperable problems of its own, and that many of these problems stem from its

commitment to another false assumption often made by writers on the problem of consciousness.

Chapters 6 and 7 expose that assumption and develop a solution to the problem of consciousness that is made available by its rejection, a solution suggested in some neglected writings of F.A. Hayek. Chapter 6 defends Hayek's view that even our knowledge of the internal world, of the mind itself, is knowledge only of structure, not of intrinsic qualities, and that a rejection of the notion of the qualitative features of consciousness as intrinsic, coupled with a recognition of their irreducibly first-person character, opens the way to a non-physicalist version of functionalism which has all the strengths of the Russellian mind-brain identity theory and none of its weaknesses. Chapter 7 examines the implications Hayekian functionalism has for the problems of intentionality and rationality, and for philosophy in general.

Some final remarks before proceeding: The literature in philosophy of mind, even just on the problem of consciousness, is vast, and I have by no means tried to survey all of it. What I say about the problem of consciousness in chapters 1-3 does try to take account of the major arguments on all sides, but I have made no attempt to track down everything that has been written. So while the objections I make to the positions I disagree with are, unless otherwise indicated, my own, it is always possible (perhaps even probable) that what I have to say echoes something written somewhere by someone else; and for the same reason, there may be existing responses to some of the objections I make which I do not consider. There is, in any case, a tremendous

amount of literature attacking both the physicalist and property dualist sides, and since I wish to call a plague down on both houses, I am happy to direct the reader to it if he should find gaps in my case. Hopefully I say enough to make plausible the view that each side at least appears to face equally insurmountable problems, so that there is motivation for trying to develop an alternative to both.

For similar reasons, I have had to make a number of assumptions in developing the argument of this essay which I do not attempt fully to defend, but which are widely enough shared by philosophers of mind of whatever persuasion that I believe the resulting lacunae are not glaring. First, I follow the common assumption that supervenience of the mental on the physical must involve there being no logically possible world in which the physical facts are the same as in ours, but in which the mental facts are different, so that the possibility of such a world would imply the falsity of physicalism.[4] Second, I assume that a broadly *materialistic* version of functionalism, even if it fails as a theory of consciousness, is adequate as an account of the propositional attitudes, namely beliefs, desires, and the like (and this plays a role in my response to property dualism).[5] This is obviously something most physicalists will

---

[4] See Chalmers 1996, chapter 2 for a detailed defense of this claim. Chalmers does argue that even if facts about consciousness fail logically to supervene on physical facts, they nevertheless *naturally* supervene on them, in the sense that given the way the natural world happens to be, facts about consciousness covary in a law-like way with physical facts; but as we'll see, precisely because they *merely* naturally supervene, Chalmers adopts property dualism rather than physicalism.

[5] Of course, beliefs, desires, and the like are often *associated* with conscious experiences (e.g. my desire for something may involve the experiencing of certain emotions), and I am of course not claiming that a materialist form of functionalism is adequate as an account of those experiences. I assume only that such a form of functionalism can account for the beliefs and desires themselves, rather than the conscious experiences associated with them (or, as it might be put instead, that it can account for the non-conscious aspects of beliefs and desires, rather than the conscious aspects).

have no problem with, and many property dualists (e.g. Chalmers) accept it too, the real problem with materialistic functionalism seeming to lie with its treatment of consciousness. (John Searle is an exception, since his famous "Chinese Room" argument is intended to undermine functionalism even about the propositional attitudes; but I respond to that argument in chapter 7, after defending a non-physicalist version of functionalism about consciousness.)

Fortunately, the Russellian and Hayekian views I am most concerned with have *not* themselves spawned an unmanageably large literature, so I believe I have been able to take account of more or less everything other writers have had to say about them. But important and even revolutionary as I take these views to be, I believe that they do *deserve* a much larger literature. If this essay accomplishes nothing else, it will at least have contributed to expanding it.

1. The qualia problem

(a) Explaining consciousness: easy and hard problems

The expressions "conscious" and "consciousness" have a number of uses in ordinary language, and more or less all of the states, processes, and events they are used to describe are of interest to the student of the mind. But the "problem of consciousness" as it has come in recent years to be understood in philosophy of mind and cognitive science involves only one aspect of the various phenomena we think of as conscious phenomena. This aspect is often referred to as the *qualitative* or *phenomenal* aspect of conscious experience, and the defining features of this aspect are referred to variously as *sensory qualities* or *phenomenal qualities*, or, most commonly these days, *qualia* (the plural form of *quale*).

What *exactly* a quale is (and, as we shall see in chapter 2, even whether such features exist) is a matter of some controversy, but the basic idea is quite intuitive. Think of the experiences you're having right now as you read this chapter. You're having certain visual experiences: the sight of black ink on white paper, and of whatever other objects are in your visual field, perhaps the brown, grainy wood of a desk, the luminous green of a desk lamp, and so forth. You are also no doubt having certain other experiences: experiences, perhaps, of the sound of a whirring fan and of distant traffic, the smell of coffee in a nearby cup, the feel of the chair beneath you, and a dull sensation of pain in your sore back. You might also be having certain thoughts: maybe you're thinking of another book you've read on the same topic, or of

the lawn you have to mow once you finish the chapter. And associated with these thoughts might be various memories, mental images, emotions, and the like. Now these conscious phenomena have a number of features: They are largely the result of events taking place outside you (e.g. light being reflected off the page, desk, etc., which strikes your retinas), and many of them involve the discrimination of objects in your environment; and they are all likely to play a part in causing certain behaviors (e.g. turning the page, picking up the cup of coffee, etc.), including verbal reports of their occurrence. But in addition to such features, there is the fact that these various conscious experiences have a certain *feel* to them, that there is, as Thomas Nagel puts it, "something it is like" to have them (1974). That is, there is, in addition to questions about the causes, effects, etc. of your experiences of the greenness of the lamp, the smell of the coffee, the pain in your back, and so on, the question of *what it is like* to have those experiences. And the features by virtue of which there is something it is like to have conscious experiences are what philosophers refer to as qualia.[1]

The most striking feature of qualia is their apparent *subjectivity* or *privacy*, the fact that they seem directly accessible only from the *first-person* point of view. The pain you feel and the mental images you might be having, to take only the two most clear-cut cases, though their existence is as evident to you as anything could be, are

[1] Though I speak here of qualia as *features* or *properties* of conscious experiences – as is generally done by writers on this subject – they are also sometimes treated by philosophers as if they were *particulars* in their own right rather than mere properties of particulars, and the distinction between properties and particulars is not always kept in mind in discussions of qualia. I will continue to treat qualia as properties, but we'll have reason later to address the issue of whether they are appropriately thought of as particulars.

not knowable by anyone else except indirectly, either as a result of your reporting them, or perhaps by inference from your behavior or the state of your nervous system. A surgeon operating on your back could observe your muscle and nerve tissues, and even any damage to them that might exist, but it seems (at least prima facie) obvious that what he won't observe is *the pain itself.* A neuroscientist examining your brain would see the various neuronal structures in all their vast complexity, and might even be able to detect a great deal of activity occurring within them, but surely he would not be able to see *the mental image itself.* And the same is true of all the other sorts of qualia we've alluded to, e.g. the greenness of the lamp, the sound of the fan, the smell of the coffee, and the feel of the chair. Obviously, other people can know what the lamp looks like, the coffee smells like, and so forth, but the point is that the particular *experiences* you are having of these things *themselves* are not accessible to anyone but you. The visual, auditory, olfactory, tactile, and gustatory sensations we have appear directly knowable only through *introspection*, "from the inside," from the first-person point of view, even though the physical objects (lamps, desks, fans, coffee, etc.) and neural processes that produce them, and the behavior that they in turn produce, are knowable from the third-person point of view, from the outside, through sensory perception.[2]

[2] The idea of subjectivity might be made even more clear by thinking of hallucinations: The experience I have when I hallucinate a chair might be qualitatively identical to the one you have when you really see a chair, and obviously, there is nothing about my experience itself which is accessible to anyone outside me, since the chair doesn't really exist. These subjective features, common to both the hallucinatory and veridical experiences, i.e. the redness of both the real and the hallucinatory chairs

Explaining qualia or the phenomenal aspect of conscious experience, that in virtue of which there is something it is like to be conscious, is, as I've said, what has come to be regarded as *the* problem of consciousness among contemporary students of the subject. Indeed, it is now, following David Chalmers (1995, 1996) commonly referred to as the "hard problem" of consciousness, a problem which poses a special difficulty for current approaches in philosophy, cognitive psychology, artificial intelligence, and neuroscience, a difficulty not posed by other features of consciousness.[3] Such problems as that of explaining an organism's ability to discriminate and react to objects in its external environment, accounting for its ability to monitor and report on many of its internal states, and even discovering the neural mechanisms that underlie various conscious experiences are said (again following Chalmers) to be "easy problems" of consciousness. That is not to say that they are trivial or even solvable without a great deal of further research; it is rather just to say that they seem fully solvable by means of further applications or extensions of current methods and theories. There seems to be no difficulty in principle in explaining, in terms of the neuroscientific and computational concepts presently available, how an organism exhibits just the behavior it does in response to internal and external stimuli; even where there are gaps in our knowledge, the problem is one of filling in details,

---

and so forth, are qualia. We will look in more detail at the importance of hallucinations and related phenomena for understanding the issues at hand when we get to chapter 4.

[3] Though Chalmers has become widely known for making the distinction between hard and easy problems of consciousness, the distinction, if not the exact terminology, is (as he would be the first to admit) actually implicit in discussions of the topic going back decades, if not centuries. And indeed,

not of having radically to overhaul our general theoretical framework. Nor does there seem to be a problem in finding out, at least in general terms, exactly what processes are occurring in the nervous system when various conscious experiences occur. But just *why* any of these physical processes are associated with conscious experiences, with qualia or a subjective, phenomenal aspect, is something else altogether; explaining *that* fact seems very difficult indeed.

The problem, in a nutshell, is this: We know that we have conscious experiences, and in particular that these experiences have a subjective, phenomenal aspect. But this seems to be a fact over and above the various physical facts about ourselves we also know. For even after we have arrived at a complete explanation of behavior and of such mental functions as discriminative ability and the like in the material and computational terms of neuroscience, cognitive science, and so forth, it seems that there remains to be explained the fact that such behavior and mental functions are accompanied by qualia. The physical facts revealed by these sciences are objective, third-person facts, while the facts about qualia are subjective, first-person facts, and the accumulation of knowledge about the former will, it appears, never yield an understanding of the latter.[4] But the methods and resources of these sciences are the only ones we have. At any rate, they seem to be the only ones consistent with the

---

even talk of a "hard" problem posed by qualia can be found pre-Chalmers in such writers as Colin McGinn (1991, p. 1) and Galen Strawson (1994, p. 93).

[4] This is why establishing a law-like empirical correlation between certain physical facts and facts about qualia would not suffice to solve the problem; for even if they are *correlated* with physical facts, facts about qualia seem nevertheless to be *further, non-physical* facts. Chalmers, as we'll see, argues

general picture of the world revealed to us by modern physics. So it is difficult to see how an explanation of qualia is possible even in principle; or in any case, if an explanation is to be had, it appears that it would have to involve a radical revision of the picture of the world that we have inherited from modern science, a rejection of the prevailing materialist or physicalist consensus.[5]

That, as I said, is the "hard problem" of consciousness in a nutshell. To make the problem (which we will from now on refer to as simply "the qualia problem") clearer, it will be necessary to look at the most influential of recent arguments alleged to show that the facts about conscious experience are inexplicable in physical terms, and thus inconsistent with materialism. The first of these is:

(b) The knowledge argument

Frank Jackson's (1982, 1991) parable of Mary the neuroscientist sets the stage for his version of the knowledge argument. Mary has spent her entire life in a black-and-white room, never having had color experiences. (We can also imagine that she has always worn a black or white suit which covers here entire body, and add any other details required to ensure that she's never seen any colors.) She has also, however, mastered neuroscience and all those parts of any other disciplines relevant to

---

for the possibility of such a law-like correlation, but precisely because the correlation would be merely empirical, he takes his position to be a kind of property dualism rather than a kind of materialism.

[5] I am well aware that some writers, most famously Roger Penrose (1989, 1994) and Michael Lockwood (1989), would deny that modern physics entails a worldview on which qualia cannot be accounted for. But their view is not really at odds with what has been said, for their view is precisely that contemporary materialists have a deficient conception of matter itself, a conception not sufficiently informed by quantum mechanics. So their position actually falls into the camp of those who argue that an explanation of qualia, if it is to be had, must involve a radical revision of the

the study of the brain and its functions; she's done so by reading all the relevant books (in black and white, of course) and seeing lectures on a black-and-white television monitor. In fact, she knows *all the physical facts there are to know* about the brain and its functions, because she lives at a time when neuroscience and all related disciplines have been completed. So Mary knows all the neurophysiological, chemical, and physical facts, as well as all the functional facts (e.g. the facts that might be uncovered by a completed cognitive science), that there are about the nervous system, including all the physical facts about color perception: she knows exactly what goes on in physical terms when someone sees a red apple, say, e.g. all the facts about light, optics, the structure and function of the eye and optic nerve, etc. etc., even though she's never seen a red apple herself. In short, if *physicalism* is true, Mary knows everything there is to know about mental phenomena, and in particular about the experience of seeing a red apple. But now suppose that Mary is finally released from the room and actually sees a red apple herself for the first time. Does she learn something new? Clearly she does: she learns *what it's like to see red*; she discovers the *qualia* associated with seeing red.

The lesson of the parable can be summed up in the following argument:

(1) Mary, before her release, knows all the *physical* facts that there are about conscious experience.

---

picture of the world that (the standard, materialist or physicalist, *interpretation* of) modern physics has given us. And Lockwood's views will be discussed at length in chapters 4 and 5.

(2) Mary, before her release, does *not* know *all* the facts that there are about conscious experience, since she learns something after her release, namely the facts about the qualia involved in seeing something red.

Therefore,

(3) The facts about qualia are facts over and above the physical facts.

And thus,

(4) Physicalism is false.

Thomas Nagel presents a related argument in his famous paper "What Is It Like to Be a Bat?" (1974). The argument hinted at by the title of his paper is just this: given that we knew all the physical facts there are about the experiences a bat has in getting around, in its unique way, by means of echolocation, there would still be something we *wouldn't* know, namely *what it's like* to have such experiences; we would lack knowledge of the *qualia* involved in echolocation, qualia which are no doubt very different from any we experience. So there are more facts than just the physical facts. So physicalism is false.[6]

---

[6] Though Jackson's and Nagel's arguments are frequently treated as two versions of more or less the same argument, Jackson, in his 1982 paper, characterizes Nagel's argument as of a different type from his own, since Nagel's presentation (unlike Jackson's) leans heavily on the *indexical* character of knowing what it's like *to be* a certain kind of creature or a certain person. Even if we *did* know all the qualia a particular person or creature has experienced, we might, Jackson suggests, still not know what it's like *to be that particular* person or creature; so that the problem about qualia that Jackson is concerned with is not, he says, quite the same as the problem Nagel is concerned with. Nevertheless, Nagel's argument is clearly partially relevant to the same issues as Jackson's, and provides an example that is in many ways more vivid than Jackson's, given the remoteness from our own qualia a bat's qualia must have. So I will follow common practice here and treat their arguments together, with the cautionary note that Nagel's argument raises further interesting issues about indexicality, issues we won't concern ourselves with here.

The upshot of both versions of the knowledge argument is that having knowledge of all the *physical* facts does not entail having knowledge of *all* the facts, and in particular, it does not entail having knowledge of the facts about qualia. Therefore, Jackson and Nagel conclude, physicalism is false; or at the very least, there is a prima facie problem of explaining how it could be true, of explaining how qualia can be accounted for in purely physical terms.

Now among replies to the knowledge argument, we can (following Robert Van Gulick, 1993) distinguish between those which grant that Mary does learn something when she leaves the room (or that we would learn something, even if we already knew all about bat neurophysiolgy, etc., if we could somehow find out what it's like to be a bat), and those which do not grant this; and these categories are, I think, more or less co-extensive with, respectively, the category of those which grant that the argument poses a strong prima facie challenge to physicalism (and insist nevertheless that that challenge can be met) and that of those which do not grant this, but which allege that the argument actually has no force at all. The former sort of reply, which I take to be more challenging, will be dealt with in chapter 3; but I will say a little about the latter sort here.

The less charitable sort of objection has been put forward by Paul Churchland (1989a, pp. 64-65) and Daniel Dennett (1991, pp. 399-401), and suggests that those who find the argument challenging have simply not taken seriously what would be involved in knowing *everything* physical that there is to know about color perception,

etc. The current gaps in our knowledge are just too great, they say, for us to be confident that we can have a sufficient grasp of what sorts of things Mary would and wouldn't know if she did have all the physical knowledge; so that for all we know, she *would* in fact know what it's like to see red, just by virtue of having *all* the information a completed neuroscience, physics, chemistry, etc., would provide.

The obvious first reply to this is that it sounds question-begging: the objection appears to be that, since all the facts there are *are* physical facts, facts about qualia *must also* be physical facts, so that Mary *would* know them even before leaving the room; and whether all facts are physical facts is precisely what's at issue. But of course, Churchland and Dennett could simply insist that it is Jackson and Nagel who are begging the question. After all, in philosophy, one man's modus ponens is frequently another's modus tollens; and debates which hinge simply on the question of which party is assuming precisely what is at issue are seldom conclusively settled (and seldom edifying). A reply which accuses Churchland and Dennett of putting forward an (always less than satisfying) argument from ignorance ("We don't know *for sure* that she *wouldn't* know, so...") is a little better; but only a little, since, again, they could conceivably turn the tables on Jackson and Nagel here as well (accusing them of arguing: "We don't know *for sure* that she *would* know, so...").

In any case, the main problem with Churchland's and Dennett's suggestion is that it simply fails to take seriously the notion that the knowledge the sciences provide us with is knowledge of *third-person* or objective facts, and that qualia are *first-person*

or subjective phenomena. When this is kept in mind, it is clear why Jackson and Nagel should think that even complete knowledge of the former sort would not yield knowledge of the latter sort. Even given the current gaps in our knowledge, the point they are making is that as long as the knowledge of physical facts yet to be discovered is of the *same general sort* as that acquired thus far, it will not amount to knowledge of the facts about qualia. Now it might be suggested that the advance of physical science could involve an *inclusion* of what we've been calling first-person facts within its purview, so that a future neuroscientist like Mary *would* know all the facts about qualia by virtue of having mastered a completed neuroscience, physics, etc. But then the future science we're talking about would be one which has either *radically revised* its (currently exclusively third-person) conception of the physical, or accepted *non-physical* features into its ontology (depending on how qualia come to be classified), so that it could no longer serve to save physicalism or materialism, as currently understood, from the knowledge argument. And in fact, such a revision of our conception of the world is exactly what many *sympathizers* with the knowledge argument have thought that argument calls for, as we'll see when we come to chapter 4.[7]

[7] In fairness, Churchland and Dennett present the objection we've been considering in the context of some other objections, which might serve as the bases for rejoinders to the reply to them I've given: Churchland also argues that even if Mary does learn something, she only comes to know in a new way facts she already knew; and Dennett argues, first, that the very notion of qualia is incoherent, and second, that that notion leads to epiphenomenalism, which, he argues, is false. I'll be dealing with these objections in chapters 3, 2, and 3, respectively.

(c) The zombie argument

The basic idea of the zombie argument is as old as the mind-body problem, really, but the "zombie" label, and the particular form of the argument it names, have only recently come to be discussed widely in the literature. The sort of zombie the argument is concerned with is not the sort familiar from the movies, which lumbers about clumsily, looking for human flesh to eat, very probably having (pleasurable?!) gustatory experiences when it does so. Rather, we are to imagine creatures which are physically, functionally, and behaviorally identical to ourselves, down to the last molecule, indistinguishable from us by any third-person means, but which nevertheless have no conscious experiences whatsoever; the physical facts about them are the same as those that hold true of us, but none of the facts about qualia which hold true of us do so for them. This sort of zombie seems perfectly possible; but if it is possible, the zombie argument concludes, physicalism must be false. David Chalmers (1996, p. 123), appealing to the idea of a logically possible world where *everyone* is a zombie, states the argument more explicitly as follows:

(1) In our world, there are conscious experiences.

(2) There is a logically possible world physically identical to ours, in which the positive facts about consciousness in our world do not hold.

(3) Therefore, facts about consciousness are further facts about our world, over and above the physical facts.

(4) So materialism is false.

The zombie argument is clearly related to the knowledge argument, and they might be seen as complementary, two ways of making the same point, namely the point that facts about qualia are facts over and above physical facts: the zombie argument makes it in *metaphysical* terms (i.e. in terms of what the *obtaining* of certain facts would and wouldn't involve), and the knowledge argument makes it in *epistemological* terms (i.e. in terms of what our *knowing* about certain facts would and wouldn't involve). A consequence of this is that (slightly different versions of) some of the same sorts of objections might be thought to apply to both.

For instance, the objection against the knowledge argument already considered clearly can be adapted for use against the zombie argument as well. For premise (2), a claim about the *possibility* of a zombie world, is typically justified (as it is by Chalmers, 1996, pp. 96-99) by reference to the *conceivability* of such a world. And perhaps Churchland and Dennett would say in this case that, just as it isn't clear (they claim) that Mary wouldn't know what it's like to see red by virtue of knowing all the physical facts, so too it isn't clear that conceiving of *all* the physical facts being just as they are in our world wouldn't involve conceiving of the facts about consciousness obtaining too. But the reply in this case would be analogous to the reply in the case of the knowledge argument: the point is that conceiving of all the physical facts would be conceiving merely of *third-person* facts, while (some of) the facts about consciousness, namely the facts about qualia, are *first-person* facts. Thus it does

indeed seem clearly conceivable that all the physical facts could obtain without all the facts about consciousness obtaining.

But another, more basic, objection can be leveled against the justification of premise (2): even if a zombie world is conceivable, why suppose that such a world is really logically possible? That is, why accept the supposition that conceivability entails possibility? Maybe it's false; in which case, premise (2) lacks any justification.

We'll consider some further objections to the zombie argument later on, but I want to try to defuse this one here, since it tries to stop the argument before it even gets going. The supposition that conceivability entails possibility, which, following William Hart (1994, p. 266), we'll call "Hume's principle," clearly has a great deal of intuitive plausibility. After all, the realm of the *logically* possible, even more than the realm of the *naturally* or *physically* possible, is on anyone's reckoning pretty vast. Though it isn't physically possible for me to fly, it is surely *logically* possible, i.e. there is no contradiction involved in the supposition that I might be able to; or, to use Chalmers's example (1996, p. 96), though a mile-high unicycle might not be physically possible, it surely seems as logically possible as a 3-foot or 300-foot unicycle does. Such examples, about which there would undoubtedly be widespread agreement, can be multiplied indefinitely. And the point is, it is hard to see on what basis anyone could accept these examples as genuine logical possibilities unless they at least implicitly accepted Hume's principle. For why does anyone judge that flying is logically possible, or that a mile-high unicycle is logically possible, if not on the basis

of the fact that he can *conceive* of these things, that he can give a *coherent description* of them? Even our judgements that certain things, round squares, say, are *not* logically possible seem to rest at least partly on our *inability* to conceive them; and should someone come along who *was* able to conceive of such things, and to present to us a coherent account of how they could, after all, exist, it is hard to believe most people wouldn't revise their judgements concerning their logical possibility accordingly. (Though I do *not*, in saying this, mean to suggest that an advocate of Hume's principle must accept the very different, and surely false, claims that inconceivability entails logical impossibility or that logical possibility entails conceivability.) One good reason for accepting Hume's principle, then, is that there seems to be no way to justify claims about logical possibility we all (including critics of the zombie argument) know (or at least believe) to be true other than by appealing to it; and in so far as we all accept such claim, it seems we all at least implicitly accept the principle anyway. Of course, this is all consistent with Hume's principle being false. But in this respect, that principle is no worse off than many things we believe and can't help believe: surely there isn't a refutation of skepticism about the external world, or induction, or the reality of the past, that is any stronger than this defense of Hume's principle; but no one would deny the rationality of believing in the external world, etc. on that account.

Another consideration that speaks in favor of Hume's principle is the enormous difficulty that must face anyone who wants to try to refute it. For refuting it would

have to involve presenting a counterexample to the principle; and it is hard to see how there could be any convincing counterexample, a counterexample which is *more* plausible than the principle itself. Such a counterexample would have to be a case of something which *is* conceivable, but is *not* logically possible. But it seems likely that anything we *thought* was conceivable and then became convinced was logically impossible would be something we'd go on to judge *wasn't* really conceivable after all. (Anyone who's introduced undergraduates to the notions of logical possibility and impossibility knows this: Often a student *thinks* he can conceive of a round square, and thus that round squares might not be impossible after all, until he's made to realize that what he's *really* conceiving of is not a round square at all, but of a *square* he's *calling* "round," or of a "round square," where the word "round" is *redefined* to mean the same as "square," or of a shape that isn't a square, but merely has three straight sides, like a square does, and one round one, which a square doesn't.)

Indeed, it may be that such a counterexample is *impossible*: for it would arguably have to be a case of something which can be given a (genuinely, and not just apparently) *coherent description*, and yet *involves a contradiction*; and the notion of such a case itself appears incoherent or contradictory. If so, this would entail that Hume's principle is a *necessary truth*! However, I won't do more here than merely suggest this as an interesting possibility, since fully to defend this claim would take us far outside the scope of this essay. Suffice it to say that, given the principle's initial

plausibility, the burden of proof is clearly on anyone who wants to reject it, and that burden is unlikely to be met.[8]

One final point: Even those who would reject Hume's principle would make a distinction between what is logically possible and what isn't. And since no one, least of all an opponent of the zombie argument, would take an "anything goes" attitude toward the question of what is to count as logically possible, the burden is on those who reject the principle to provide some *alternative* criterion for determining what is logically possible. Moreover, if the zombie argument is to be undermined, it would have to be a criterion which does not justify premise (2), as Hume's principle does. But since no such alternative criterion has been offered, it is hard to see what grounds there could be for rejecting premise (2) even if Hume's principle were shown to be false.

(d) Other arguments: modal, inverted spectrum, Chinese nation, explanatory gap

The knowledge and zombie arguments are, in my view, the ones that really get to the nub of the qualia problem, which is, as I've indicated, the gap between the third-person or objective facts about the physical world, and the first-person or subjective facts about qualia. There are a number of other influential arguments for roughly the

---

[8] It might be thought that (what are, since Saul Kripke's influential 1972, widely recognized to be) *necessary* truths such as "water = H2O" are nevertheless *conceivably* false, and thus provide counterexamples to Hume's principle. But to assume this would be to fail to take account of the subtleties of the semantic situation. Roughly, the statement only *seems* conceivably false when one fails to keep in mind that "water" and "H2O" are rigid designators, i.e. they name the *same* substances in *all* possible worlds; when one keeps this in mind, the illusion of conceivability disappears: the falsity of "water = H2O" comes to seem as inconceivable as the falsity of "water = water." See Chalmers (1996, pp. 52-71) for a detailed discussion of these semantic issues (and a

same conclusion, but in my view, they are either less straightforward or fail to get to the heart of the matter. Nevertheless, because they are so influential, and also at least bolster and perhaps further clarify the point, we would do well to consider them briefly.

Saul Kripke's groundbreaking work in semantics (1972) led, among other things, to a unique argument, sometimes called the *modal argument*, against the mind-brain identity theory. That theory, as developed by such writers as J.J.C. Smart (1959), presented the purported identity between brain states and mental states as a contingent identity, an identity of the sort such identities as that between water and H2O, or between lightning and electrical discharges, were thought to be. But as Kripke argued, identities of the latter sort are in fact *necessary* identities, identities which hold true in *every possible world*. For e.g. "water" and "H2O" are *rigid designators*, that is, they name the same things in every possible world, so that "water = H2O," if true in any possible world, is true in all, that is, is a necessary truth. But mentalistic terms such as "pain," and terms for kinds of brain states such as, say, "the firing of C-fibers," are also rigid designators; so that if, say, "pain = the firing of C-fibers" is true in any possible world (such as ours) it must be true in all. But clearly it *isn't* true in every possible world, Kripke argues; for we can conceive of worlds in which pain exists without the firing of C-fibers, or in which the firing of C-fibers

---

defense of the principle that conceivability entails possibility against the objection under discussion at pages 67-68 and 98).

occurs in the absence of pain. So "pain = the firing of C-fibers" is false, and thus so is the mind-brain identity theory.

There is an obvious similarity between this argument and the zombie argument, namely the appeal to the possibility of a world where the physical facts, such as the facts about C-fibers, are the same, and yet the facts about qualia do not obtain. But Kripke's argument appeals to some complicated and controversial positions in the philosophy of language which make his argument much less straightforward than the zombie argument, and which we cannot get into here. So for the purposes of this essay, we will assume that the qualia problem can be adequately spelled out without appealing to Kripkean semantics, and thus without appealing to the modal argument.[9]

The idea of the *inverted spectrum* has a long history in philosophy, going back at least to Locke. It goes like this: it seems possible that another person, even one who is physically, functionally, and behaviorally identical to you, could have color experiences which are inverted relative to your own; that is, what you see when you look at what you both call red, for instance, is what he sees when he looks at what you both call green, and vice versa, and this difference would nevertheless not register in what either of you said about red and green objects or how you interacted with them. And the lesson often drawn from this is that certain facts about qualia, e.g. facts about the nature of the experiences we have when we look at red and green objects, are facts

[9] See Chalmers (1996, pp. 146-149) for a detailed discussion of the similarities and differences between Kripke's argument and the zombie argument, and of some weaknesses Kripke's argument arguably exhibits in those respects in which it differs from the zombie argument.

over and above the physical facts, so that it is false to say that the neurophysiological, chemical, etc. facts about us are all the facts there are.

This argument does without a doubt provide a vivid way of presenting the qualia problem, and may even, for some readers, have more intuitive force than the knowledge or zombie arguments. Still, it does not, I think, get quite as close to the heart of the matter as those arguments do, because it leaves it open that the physical facts *might* still entail at least the *existence* of *some* qualia or other, and just fail to determine their precise *nature*. The *complete* break that seems to exist between the third-person facts and the first-person facts is clearer from the other arguments.

Ned Block's (1978) *Chinese nation* argument, though also vivid, is similarly less clear in getting across the completeness of the break. This argument was presented as a criticism of one specific version of materialism, namely (the standard, physicalistic version of) *functionalism*, the view that what makes a given type of mental state the exact type it is is its functional role, that is, the causal relations it bears to environmental stimuli, other mental states, and behavior. So what makes a pain a pain, on this view, is just that typically it is caused by bodily injury, causes other mental states such as distress, and either directly or together with other mental states causes such behaviors as wincing, crying, and nursing of the damaged area of the body; *any* state of *any* system, whether composed of neural tissue, silicon chips, or whatever, that played *just* this role would, by virtue of doing so, *just be* a pain. Block's objection to this was that there are all sorts of systems that could conceivably

have states playing just the causal roles played by pain and other mental states, and yet would nevertheless clearly lack those mental states themselves. For instance, it is conceivable that the entire population of China could be so organized that each individual plays a role analogous to that played by some portion of the brain underlying some mental state, the entire system resulting in behavior by means of connections to a robotic body; and yet it seems absurd to suppose that this vast "brain" would thereby have pains, visual experiences, or any qualia at all. Therefore, there must be more to the having of qualia than merely instantiating states which play certain functional roles.

Challenging as this argument is to the functionalist version of materialism, it nevertheless fails, by itself, to get across the idea that facts about qualia seem to be facts over and above *any* set of physical facts about *any* physical system, *including* the brain, not just unusual systems which mirror the functional organization of the brain. So like the inverted spectrum argument, it seems to me less fundamental to the qualia problem than the knowledge and zombie arguments are.[10]

We should also note, finally, Joseph Levine's *explanatory gap* argument (presented in, among many other places, his 1993). Levine holds that arguments like

[10] It might be suggested that the Chinese nation argument does play an indispensable role in fleshing out the qualia problem in that it makes explicit the idea that it is not just physical systems per se which can come apart from qualia, but even physical systems qua having a certain functional organization. As I will try to show later, however, the intuitive plausibility of the argument really turns out to rest after all on the same intuition underlying the zombie argument, namely the intuition that a physical system as such can lack qualia; despite appearances, it has no tendency to show that a *functional organization* like ours can come apart from qualia. To spell out why this is so, however,

those of Jackson and Kripke show physicalism to be inadequate in that they demonstrate that qualia cannot be *explained* in physical terms, since there is no entailment from physical facts to phenomenal facts. The twist in Levine's position, however, is that he holds that this is only an *epistemological* result, not a *metaphysical* one; the explanatory gap between physical facts and facts about qualia is only a gap in our knowledge, not a gap in reality between two fundamentally different kinds of fact. It's not that facts about qualia are not physical facts; it's rather that we don't *understand how* they are. So Levine's view is much less bold a challenge to physicalism than the other arguments we've looked at.

The problem with Levine's position is that it rests on grounds we've already found to be suspect. He rejects Kripke's argument as having any *metaphysical* force because, he says, there's no good reason to accept what we've been calling Hume's principle that conceivability entails possibility, a principle Kripke's argument clearly appeals to; that argument, he claims, shows only that we don't *understand* how qualia can be identical with brain states, not that they *aren't* in fact identical. But as we've seen, such a dismissal of Hume's principle will not do.[11] He denies any metaphysical force to Jackson's argument because he says it could be that what Mary learns upon

---

would require spilling the beans already on my own solution to the qualia problem, and that must wait until chapter 6.

[11] Levine makes much of the fact that the way things seem, however clear and distinct, isn't a sure guide to the way they are (1993, p. 123), so that something that seems conceivable may not really be possible. But this is no better a reason to reject Hume's principle that the fact that our perceptions, however clear and distinct, are sometimes misleading, is a reason to reject the principle that perception is a reliable guide to reality. In fact, while there are counterexamples to latter principle, there are, as noted earlier, *none* to the former. So perhaps Hume's principle is in even better shape

leaving the room *may* after all be "physical information" in the sense of "information about a physical event or process," which, he says, is the only sense in which "any reasonable physicalist is committed to the claim that all information is physical information" (1993, p. 125). If so, there would be an epistemological gap between knowledge of physical facts and knowledge of qualia (since Mary has information about the former but not the latter while in the room) but not a metaphysical gap (since it might still be that the information she lacks while in the room is just further information about physical events and processes). The problem with this is that the conception of "physicalism" Levine seems to be left with is so loose that it would be compatible with *property dualism*: even if the latter were true, all information would still be "information *about* a physical event or process," because on the property dualist's view (unlike the view of the substance dualist), mental events and processes are themselves identical with brain events and processes, and merely have non-physical *properties*. Even a fact such as the fact that the firing of C-fibers is associated with a non-physical quale would be "a fact about a physical event or process." In short, Levine's suggestion seems little different from the suggestion, considered above, that first-person, subjective facts simply be *added* to our conception of the physical; a suggestion which, whatever its merits (and it has many, as we shall see later on), hardly counts as a defense of "physicalism" as usually conceived.[12]

---

than perception is, and no one would recommend abandoning the latter. (See Yablo (1993) for a detailed discussion of these issues.)

[12] Levine says, with regard to his suggested construal of "physical information": "Actually, I think any interesting doctrine of physicalism is committed to more than this, though it's difficult to pin

For these reasons, Levine's argument, though influential, seems to me not to add to our understanding of the qualia problem.

(e) The nature of qualia and of the qualia problem

The qualia problem, then, is the problem of explaining the existence of features of consciousness which, unlike anything else in the world, and certainly anything else studied or discovered by the natural sciences, appear to be essentially private, subjective, accessible only from the first-person point of view; of explaining exactly how these features relate to the rest of the world, the world of public, objective, third-person objects, events, and processes. Now I say here, as I've frequently said earlier, that these features "appear" to be essentially subjective, etc., and I've done so deliberately, so as not to beg any questions. For many solutions to the qualia problem, though they grant that there *appear* to be such features, grant *only* this, and seek to show either that such features don't truly exist at all, or that they are, after all, actually as objective and accessible from the third-person view as any others. This, at any rate, is what physicalist or materialist solutions to the qualia problem do, solutions we'll be examining in chapters 2 and 3. And in one respect, physicalist solutions to the problem would be the only solutions there could be, because the problem itself is, in a sense, only a problem for physicalism (which is certainly the impression given by the

---

down exactly how much more" (1993, p. 125). This seems to me exactly right; however, he goes on to add: "At any rate, it doesn't affect the present point," which, as what I've said indicates, I think is false. But perhaps what Levine is getting at in the argument we've been considering is the somewhat different objection that Mary, when she leaves the room, learns *in a new way* facts she *already* knew while in the room (in which case it would seem to be misleading to speak of her as acquiring

knowledge argument, zombie argument, etc.). For only physicalism denies that there are any facts over and above the physical facts. Dualistic views, for example, are hardly going to see the existence of qualia as any more problematic than the existence of matter itself.

But of course, there is a sense in which the qualia problem is a problem for any view in the philosophy of mind, at least in so far as any view is going to want to explain the exact *relationship* between qualia and the rest of the world; and in that sense, there are non-physicalistic solutions to the problem. Property dualism, which we'll look at in chapter 3, is one of these. The Russellian identity theory and Hayekian functionalism, the subjects of chapters 4 and 6 respectively, are also solutions to the qualia problem in this sense – solutions which, as we'll see, in a sense combine the solutions of both physicalists *and* property dualists.

Before we turn to examining these solutions to the problem, however, we need to say a little more about the precise nature of qualia. I've defined them as being those features of consciousness which (i) are that by virtue of which there's something it's like to be conscious, and (ii) appear essentially private, subjective, accessible only from the first-person point of view. And the first thing we want to ask is what the relationship is between (i) and (ii): are they the *same* feature?

There is a good prima facie reason for thinking they aren't, namely that no one denies that there exist properties of sort (i), but many people, particularly physicalists,

---

information when she leaves the room). In chapter 3, we'll consider the version of this objection

deny that there are any properties of sort (ii). On the other hand, when we ask what *exactly* it is that physicalist accounts are alleged to leave out when they leave out "what it's like" to have conscious experiences, it seems hard to answer without going on to say that what they leave out are just the *subjective, first-person* features of consciousness, that is, properties of sort (ii). So there does *seem* to be a link between (i) and (ii); and hopefully by the time we've looked carefully at the positions explored in chapters 4-6, it will have been shown that there *is* an essential link between them, and in fact that (ii) is the more basic feature.

But the reader might be wondering: Haven't I still left something out? Aren't there a number of other features which are essential to qualia? It is often thought that there are. As Dennett writes:

> So, to summarize the tradition, qualia are supposed to be properties of a subject's mental states that are
>
> (1) ineffable
>
> (2) intrinsic
>
> (3) private
>
> (4) directly or immediately apprehensible in consciousness (1993, p. 385)

Privacy, as the term is used in discussions of qualia, is, of course, is just another name for the subjectivity or first-person character of qualia. But what of ineffability,

---

presented by Churchland.

intrinsicality, and direct apprehensibility? Are they not further, distinct properties of qualia?

My view is that all three of these other properties are, to the extent that they are genuine, or genuinely distinct, properties of qualia at all, less fundamental than the property of privacy or subjectivity, which is in my view *the* defining feature of a quale.[13] This is most obvious in the case of ineffability: clearly the reason qualia are thought to be ineffable is that our language typically is used to communicate thoughts about objective, public phenomena, and words are indeed typically learned by reference to such phenomena, so that communicating thoughts about what appear to be private, subjective phenomena is difficult, even seemingly impossible. So to the extent that qualia are ineffable, this is only as a consequence of their being private or subjective.[14]

Something similar can be said about direct apprehensibility. What is meant by saying that qualia are directly apprehensible is that we can know about them without having to *infer* their existence; and this is usually conjoined with the thesis that the public, objective, third-person objects and events that cause us to have qualia can themselves only be known indirectly, through inference from the existence of the

---

[13] It might seem that propositional attitudes like beliefs, desires, and the like are also private, though they aren't qualia. But strictly speaking, it is surely the *qualia* associated with, or qualitative or phenomenal features of, such propositional attitudes that are private, rather then the propositional attitudes *per se* – after all, such other features of propositional attitudes as the behavioral dispositions they are associated with are *not* private.

[14] It might be thought that if qualia are private, they must be ineffable in the strong sense that communication about them *really would* be impossible, in which case talk about qualia would be

qualia themselves. In other words, direct apprehensibility is tied to the doctrine, which goes back to Descartes and classical empiricism, that the existence and nature of the public, objective, third-person world can be known only through the "veil of perceptions" which constitutes the private, subjective, first-person world. So privacy or subjectivity appears to be more fundamental than direct apprehensibility as well: nothing that is public or objective could be directly apprehensible, since public, objective phenomena can be known, on the veil of perceptions view, only through what is private and subjective. Of course, on a direct realist theory of perception, objective, public objects *are* directly perceived. But no direct realist thinks this justifies ascribing a surprising, unique property of "directly apprehensibity" to physical objects, which just bolsters my claim that "direct apprehensibility," as the term is used in discussions of qualia, has a special sense which derives from its association with the more fundamental idea of privacy or subjectivity. (In any case, we'll be looking at all these issues in greater detail when we come to discuss perception in chapter 4.)

Intrinsicality, of all the features ascribed to qualia, is the one most plausibly regarded as distinct from and independent of privacy or subjectivity. Intrinsic properties are defined by contrast with relational properties, properties defined in terms of their relations: think of the way mental states are defined by functionalism, i.e. in terms of their relations to inputs, other mental states, and outputs. (More accurately, *the property of being in a particular mental state* is relational, on the

---

literally meaningless. This is, more or less, the view of Wittgenstein, which we will examine in

functionalist view, because a system has it only by virtue of being impacted by certain environmental stimuli, having certain other mental states, and producing certain behavioral outputs.) Intrinsic properties are just properties that aren't like this, that aren't relational. And what is there in this that demands explication in terms of privacy or subjectivity? Well, it's because qualia are thought not to be analyzable in terms of environmental inputs, behavioral outputs, and other mental states (conceived of in *purely functional* terms) – all *third-person, objective phenomena* – that functionalism is thought to "leave out qualia." And in general, the idea that qualia are intrinsic seems to arise from the sense that they are indefinable in terms of relations between *physical* objects, events, and processes. So my suspicion is that once again, the problem seems to be one of objective, third-person facts failing to yield subjective, first-person facts, in this case, a problem of qualia being indefinable in terms of objective, third-person relations. Even if intrinsicality per se can be conceived of quite independently of subjectivity or privacy, then, I suggest that the *reason* it is ascribed to qualia has to do entirely with the idea that qualia can't be shown to be an objective, public relational property.

In any case, I hope to show that qualia are in fact *not* intrinsic properties; and in particular that they are relational, but *nevertheless*, at the same time, *subjective, first-person* properties. My reasons for making this claim will have to wait until chapter 6. But hopefully it is clear already why a solution to the qualia problem which

---

chapter 2.

rejects the intrinsicality ascribed to qualia by property dualists and Russellians need by no means amount to physicalism which, unlike the Hayekian position I will be defending, rejects the notion of the mind as a subjective, first-person realm.

Before arguing for that position, though, I want first to show what is wrong with rival solutions to the qualia problem. It is to that task that I now turn.

2. Attempts to dissolve the qualia problem

(a) Wittgenstein's private language argument

Some forty years of exegesis have, unfortunately, failed to remove entirely the air of obscurity that surrounds Wittgenstein's famous argument, but that it is intended to show that there is no genuine problem about the character of sensory experience – that the problem of qualia is a pseudo-problem – is clear. The thrust of the argument is that the facts about the language we use to talk about our sensory experiences precludes there being any features of those experiences which are not accessible from the third-person point of view, that is, any features of the sort philosophers like Jackson, Nagel, and Chalmers take to be inexplicable on a standard materialist view.

Controversy surrounds the matter of the correct interpretation of the argument, which is spread out over sections 243-315 of Wittgenstein's Philosophical Investigations; controversy I cannot get into here, much less hope to settle. I will simply assess what I take to be that construal of the argument which poses the strongest challenge to the view that there is a problem about qualia. (Wittgenstein, of course, does not speak of "qualia," but rather of the idea of a private object of which one is aware when one has a sensation such as a pain. But in denying the existence of such private objects, he is denying the existence of what are today generally referred to as "qualia.")

Part of the problem of interpreting Wittgenstein here is that, as he presents his position, there seem to be a number of lines of criticism of the idea that any features of

the sort commonly called qualia exist, rather than a single argument. Nevertheless, I think the various points can usefully and accurately be summarized in the following dilemma argument:

(1) If there are qualia, then there can be meaningful discourse about them.

(2) There can be meaningful discourse about qualia only if there can be such discourse in either (i) a public language, or (ii) a private language.

(3) There can be no meaningful discourse about qualia in a public language.

(4) There can be no meaningful discourse about qualia in a private language.

(5) So there can be no meaningful discourse about qualia at all.

(6) So there are no qualia.

Premise (1) might seem to smack of verificationism, a doctrine which has taken some knocks in recent philosophical discussion; but it isn't clear that it can only be understood in that way. There's nothing in (1) which implies that for something to count as real, there has, in principle at least, to be *evidence* for or against it; it implies only that there has to be some way for us to *talk* about its existing (whether or not it really does exist), or, what amounts more or less to the same thing, some way for us to *make sense* of the claim that it exists. This sounds pretty innocuous, but perhaps even it will seem to some to be problematic, because it may seem too anthropomorphic: After all, why suppose that for something to exist, *we* have to be able to make sense of it? But the basic idea can be restated to get around this problem. The point is that, at the very least, if *we're* to be justified in claiming that something – in this case qualia –

exists, then *we* have to be able meaningfully to speak about it. In any case, I'm not going to try to challenge premise (1). Nor will I challenge (2), which seems indeed to exhaust all the possibilities. Premise (3) or premise (4), then, will have to be challenged if Wittgenstein's case is to be rebutted.

Wittgenstein's case for (3) appears to be as follows: If qualia are supposed to be essentially private, directly accessible only to the person who has them, then there is no way in principle for a person to know that what others refer to when, say, they talk about pain is the same sort of thing *he* refers to when *he* talks about it (1953, sec. 293). But this means that in a public language, a language which allows for communication between language users, there can be no meaningful discourse – no communication – about qualia, for "if language is to be a means of communication there must be agreement not only in definitions but also (queer as this may sound) in judgments" (1953, sec. 242); and if there is no way in principle to know whether other people mean the same thing I do when they talk about pain, there is no way in principle to know whether we agree in judgments about pain.

The only way we could agree in judgments about pain and other sensory experiences is if there were some evidence to which we all had access and thus to which we all could appeal – behavior, say. And indeed, for Wittgenstein, it is just this which enables us to agree in judgments; for the expressions we use to talk about sensory experiences are "connected with" or "tied up with" behavior (1953, secs. 244, 256). Any private aspect to the sensory experience thus drops out as irrelevant to

meaningful discourse about it (which is the point of the famous "beetle in the box" example, 1953, sec. 293).

But even if, for communication about it to be possible, "an 'inner process' stands in need of outward criteria" (1953, sec. 580), it might appear that a private aspect of the inner process – the quale – nevertheless exists, even if it is in principle ineffable; perhaps I can make sense of it, *to myself* at least, by means of a private language, a language which, in principle, only I am capable of understanding. Premise (4) denies this, and his argument for it is what gives Wittgenstein's "private language argument" its name. Now that argument gets us into the issue of the relationship between speaking a language and rule-following, and the idea that linguistic behavior essentially involves adherence (even if only *tacit* adherence) to a set of criteria differentiating correct from incorrect usage – a very complicated business indeed, as anyone even dimly aware of the Wittgenstein literature knows. Moreover, not only is Wittgenstein's own exposition of the argument notoriously obscure – typically so, of course, given his peculiar style – but even commentators on it generally do not try to set it out in explicit step-by-step form. Still, I think the gist of the argument can be set out fairly clearly, if at length, as follows:

A. Use of a language necessarily involves the following of rules.

B. So use in a private language of a linguistic expression 'Q' to refer to a quale must involve the following of a rule.

C. One can follow a rule only if one can in principle distinguish between correct and incorrect usage.

D. So one can follow a rule in using 'Q' only if one can in principle distinguish between correct and incorrect usage of 'Q'.

E. Now in a private language, correct usage of 'Q' could only be determined by reference to a private ostensive definition of 'Q' as referring to a particular quale.

F. So in a private language, if correct usage of 'Q' can be distinguished from incorrect usage at all, it can only be by reference to *memory* of the private ostensive definition.

G. But a quale is something accessible only to the person who has it.

H. So a private ostensive definition of 'Q' as referring to a given quale could be "witnessed" only by the person who has the quale.

I. So there would be no way for a person using 'Q' in a private language to verify his memory of the private ostensive definition – no way to distinguish *remembering* it from only *seeming to remember* it.

J. So memory could not in fact serve as a criterion for distinguishing correct from incorrect usage of 'Q' in a private language.

K. So nothing could serve as a criterion to distinguish correct from incorrect usage of 'Q' in a private language.

L. So there is no way to follow a rule in using 'Q' in a private language.

M. So there can be no use in a private language of any expression 'Q' to refer to a quale.

And from M it follows that

(4) There can be no meaningful discourse about qualia in a private language.

Now I am not in fact going to try to argue against (4), since I think Wittgenstein's basic insight that language – including the language we use to talk about our sensory experiences – is essentially social is correct (though I realize this

rather vague statement would need to be spelled out before it can be claimed to say anything interesting); not to mention the fact that the issue of rule-following is a hornet's nest that I don't wish to brave just now. Still, I do have a quibble with his case for (4), namely that I don't think it's as obvious as he implies that the lack of an objective check on memory is a special problem for the idea of a private language for qualia. Wittgenstein's claim is that I couldn't know whether or not I'm using 'Q' to refer to the same sort of quale I referred to when I introduced 'Q' by private ostensive definition, because I (obviously) couldn't ask someone if this quale is of the same sort as that, and I couldn't check my memory by other means such as by looking at a mental chart which matched 'Q' with a particular quale, since I couldn't be sure that I was remembering the chart right either. But it is easy to imagine the same sort of problem popping up in the case of reference to publicly accessible objects in a public language. Suppose I'm stranded on a desert isle or am the only person left after a nuclear war, or whatever. For all I know, I may end up misremembering the way words were used before my solitude. Does that mean I can no longer count myself as meaningfully using the expression "coconut," say, since there's no longer a way of checking the correctness of my usage (especially if we also suppose that all books with labeled pictures of coconuts have been lost or destroyed, etc.)? Surely not.

Of course, Wittgenstein would no doubt respond that the important point is that there is no way *in principle* of checking one's memory in the case of qualia, while there is in the case of public objects. But I fail to see how this is a difference that

makes a difference. Either memories are faulty or they aren't, whatever the memories are about. So if I can't *in fact* check my memory, it would seem to follow, if Wittgenstein is right, that I can't meaningfully use words, whatever the words are about. Indeed, the point seems to hold even if we don't think in terms of Robinson Crusoe type cases: Maybe *all* of us, right now, are misremembering the way we used words an hour ago; and perhaps whenever we try to check our memories by appealing to dictionaries, reference works, and the like, we misremember what we saw in the books moments after we look at them. If so, then we can't meaningfully be using expressions for public objects any more than we can for private objects. Now a response to such skepticism about memory might be that in the absence of *specific grounds for doubting* one's memory, or all of our memories, we needn't worry about whether it is trustworthy, in which case we needn't worry about whether we can meaningfully use language – indeed, this would be a very Wittgensteinian response. But this response also serves as a response to the problem Wittgenstein tries to pose for reference to qualia in a private language: In the absence of specific grounds for doubting my memory of a private ostensive definition, why suppose that I can't trust it, and thus can't meaningfully be using 'Q'? Moreover, why can't the friend of qualia simply say: "Look, I certainly remember, as well as I can remember anything, the qualia associated with the nail-gun accident I was in back when I worked construction – hell, I wish could forget them! So don't tell me there's a problem about remembering private ostensive definitions of qualia terms!" Wittgenstein would no

doubt insist that the vivid memories the friend of qualia has are *not* memories of qualia; but I don't see how he can *argue* that they are not without begging the question.

In short, the problem Wittgenstein poses for meaningful talk about qualia in a private language is simply a problem about whether one can be justified in claiming to remember something; and as such, it is not a *special* problem for a private language about qualia, not a problem that doesn't apply also to talk about public objects in a public language. It's nothing but a special case of general skepticism about memory, which, if it is a genuine problem at all, is a problem for *any* position.

Again, though, even if Wittgenstein's case for (4) fails, I am not interested here in determining whether (4) is in fact true. What I want to reject is premise (3). And the first thing I want to suggest against it is that if it were true, it would follow that behaviorism is true – in particular, that form of the doctrine known as "*logical* behaviorism." But logical behaviorism *isn't* true, and thus neither is (3).

The charge of behaviorism is, of course, an old one, and one that followers of Wittgenstein no doubt find tiresome and unfair. After all, Wittgenstein doesn't claim to be a behaviorist, and in fact *denies* that he is one (1953, secs. 304, 307). Nevertheless, I think the charge is just. If meaningful discourse about sensory experiences *is* "tied up with" behavior associated with those experiences, but *not* with anything private (which, even if it exists, "drops out of consideration as irrelevant" (1953, sec. 293)), it is hard to see how to avoid the conclusion that language about

sensory experiences *just is* language about behavior and dispositions to behavior – which is precisely the thesis of logical behaviorism.

Now followers of Wittgenstein claim that there is an important difference between Wittgenstein and the logical behaviorists, which has to do with their respective conceptions of behavior itself. "While behaviorism rejects the Cartesian picture of the mind as a private mental theatre, it accepts the attendant conception of the body as a mere mechanism, and of human behaviour as 'colourless' physical movements" (Glock, 1996, p. 57). The idea seems to be that logical behaviorism accepts the Cartesian claim that mind and body are conceptually distinct, rejects the existence of mind, and then goes on to *redefine* mentalistic expressions in terms of behavior, understood mechanistically (in a kind of Carnapian "rational reconstruction" of the mental); while Wittgenstein, by contrast, recognizes a conceptual link between mind and behavior from the start, denying that the one can be described apart from the other.

I don't wish to downplay the reality or importance of this distinction. Wittgenstein's approach clearly is in many ways different from (and more sophisticated than) that of logical behaviorism, both in terms of starting points and execution: Wittgenstein never claimed, for example, that an analysis of mentalistic expressions in behaviorial terms could ever be carried out in such a way that the latter could *replace* the former, as logical behaviorism implied (Glock, ibid.). Still, they end up, it seems to me, in much the same place: Whether you take yourself to be *redefining* mentalistic

expressions in terms of (otherwise) "colorless" behavior or rather to be finding the mental "color" to have been there all along, the end result is the same, with mental talk turning out to be nothing more than talk about behavior and dispositions to behavior.

To be sure, Wittgenstein tried to avoid this result. He sought, not to give a behavioristic analysis of mentalistic expressions, but rather to provide a middle way between Cartesianism and behaviorism: The mental is not reducible to the behaviorial, on his account, but neither can it be conceived of apart from the latter (Glock, 1996, pp. 57-58; Budd, 1989, pp. 18-19). That is, there is less conceptual "slack" between mind and behavior than Descartes supposed, but more than logical behaviorism supposes.

Intentions notwithstanding, however, this attempt at a middle way cannot work. Wittgenstein's position here strikes me as analogous to the position taken by many philosophers of religion who have been influenced by Wittgenstein, and it fails for the same reasons. Let me briefly say a little bit about that position, for I think the comparison with the philosophy of mind case is instructive. Traditionally, religious expressions such as "God" have been taken (purportedly) to refer to metaphysical entities. But a few mid-twentieth-century philosophers, convinced that logical positivism had destroyed metaphysics but still wanting to preserve some sort of religious belief, proposed *redefining* religious expressions in moral terms: "I believe in God," for example, should now mean, not that the speaker affirms the existence of some metaphysical entity, but rather that he is determined to live a moral life, or

whatever. R.B. Braithwaite (1964) was a famous advocate of this approach – an approach which, as perhaps goes without saying, was found wanting by religious believers and non-believers alike as failing utterly to do justice to what is essential to religious belief. Now certain followers of Wittgenstein, D.Z. Phillips being the best-known of them, have put forward an analysis of religious language that is often accused of being no better than Braithwaite's, seeing as it denies any metaphysical content to religious expressions, regarding them instead as expressive of a moral outlook: Talk about life after death, for example, isn't taken to be talk about continued existence beyond the grave, but rather as expressive of a certain attitude taken toward this life (Phillips, 1970). But these Wittgensteinians deny the charge: They aren't *re*defining religious expressions, they say, but only calling attention to the meaning it really had all along.

Attempts to defend this sort of view, though, inevitably seem to involve explaining the obscure in terms of the more obscure. Phillips rejects metaphysical construals of religious language: "It is a grammatical confusion," he says, "to regard this language as referential or descriptive [of some metaphysical realm]. It is an expression of value" (1976, p. 147). But he denies that his account thereby *reduces* religious language to moral language in the way Braithwaite's does. Religious language is sui generis: "If one asks what it says, the answer is that it says itself" (ibid.) It might not be too unfair to say that on Phillips' view, religious language is just like moral language... except in the respects in which it isn't. As J.L. Mackie writes:

> Phillips swings from one alternative to the other, wrapping both in obscurity, because he is seeking, but cannot find, a view that is different from both. What *he* wants to say cannot, indeed, be said; but this is a symptom not of depth but of incoherence. (Mackie, 1982, p. 226)

We can, I think, be forgiven for suspecting that something similar can be said about Wittgenstein's attempt to find a middle way between behaviorism and Cartesianism when we consider such passages as the following:

> "But you will surely admit that there is a difference between pain-behavior accompanied by pain and pain-behavior without any pain?" – Admit it? What greater difference could there be? – "And yet you again and again reach the conclusion that the sensation itself is a *nothing*." – Not at all. It is not a *something*, but not a *nothing* either! (1953, sec. 304)

> "Are you not really a behaviourist in disguise? Aren't you at bottom really saying that everything except human behavior is a fiction?" – If I do speak of a fiction, then it is of a *grammatical* fiction. (1953, sec. 307)

If we take a sensation of pain to be a *particular kind* of behavioral disposition – say, a disposition not associated with another behavioral disposition to say things like "I was only faking it!" when a friend asks about the pain, etc., then it's clear what Wittgenstein means when he says the sensation isn't a "nothing" even though it isn't a "something," i.e. a private object of introspective awareness or a quale. Similarly, if

we take the "grammatical fiction" he rejects to be the idea that the "grammar" of pain has anything to do with qualia, but is instead to be understood purely in terms of behavior, his meaning is clear. But to interpret his remarks in this way would be to interpret them as assertions of behaviorism, which he says he disavows. Yet if we *don't* give them this sort of reading, these statements seem utterly mystifying. Like Phillips, Wittgenstein appears capable of avoiding reductionism only at the cost of intelligibility.

A response the defender of Wittgenstein might give to this would be to say that his remarks only seem mystifying if we assume that talk about sensations has to be referential, so that if it doesn't refer to qualia, behavior seems the only alternative. But Wittgenstein would deny that it is referential: In the first-person case, at least, when a person uses an expression like "pain," he is not using it to *refer to* or *describe* an inner process, but rather as an *expression* of the pain, much like a groan or cry is an expression of pain (1967, sec. 472). It is hard to see how this helps defend Wittgenstein against the charge of behaviorism, though. If we take seriously the idea of "I am in pain," for example, being an *expression of* pain rather than an assertion, there's still the question of *what* the pain *is,* of which it is an expression, and we're back where we started. If, on the other hand, we don't take the "of" in "expression of" seriously, but say instead that the expression itself, a particular kind of behavior, isn't *of* the pain but just *is* the pain, then we've just given a behavioristic analysis of pain.

Moreover, Wittgenstein's view that first-person use of sensation terms is (always) expressive rather than descriptive and referential is highly implausible. If "The pain is unbearable!" is expressive and not descriptive, then surely it must be non-cognitive – without a truth value. But in that case, someone who replied "Oh come on, it's not that bad!" would not be contradicting the first speaker – which he clearly *would* be. Indeed, he *couldn't* contradict him, any more than one could contradict an animal's cry; for one can contradict someone only when he assigns a different truth value to a claim made by the other person. Or are we to suppose that "I am in pain!" and its like are expressive *except* in those cases where one is lying? In those cases, I suppose, it would be descriptive (and thus cognitive) but always false – and "It's not that bad!" would be true (though, in cases of a sincere expression of pain, it would be... what? False? Neither true nor false?). Very odd!

Now Wittgenstein does hold that *third-person* cases – such as "He's in pain" – are descriptive; an eminently plausible view to take. But then we have the strange consequence that "pain" has one use when I apply it to you and another when I apply it to myself. (Surely this smacks of just the sort of bifurcation in meaning Wittgenstein condemns the Cartesian view for threatening us with, in that that view appears to open the way to "pain" meaning a private object or quale when I apply it to myself, but to behavior when I apply it to you – since I can't, in principle, know anything about your qualia.) So "I have a dollar" and "He has a dollar" are of the same logical type, but "I'm in pain" and "He's in pain" are not? Is this really more plausible than the view

Wittgenstein is criticizing? After all, we can certainly think of cases (and actual cases of usage are just the sort of thing Wittgenstein urges us to look at) where "I'm in pain" *seems* descriptive: A doctor comes into the waiting room and asks "Who here is in pain?" "I'm in pain. He's in pain, too." Surely the first sentence is as informative and descriptive as the second! (And what, on the Wittgensteinian account, could we possibly make of "We're in pain!" spoken by one person about himself and another? Is it partly descriptive (and thus true) and partly not (and thus partly neither true nor false)?!)

Support for Wittgenstein's position might nevertheless be sought in his famous related view that it makes no sense to claim to *know* that one is in pain – that such a claim is without significance or literally meaningless, in that the concept of knowledge simply doesn't apply here (1953, secs. 246, 288, 408). Perhaps one could argue this way: If "I am in pain!" were descriptive, the sort of thing that could be true or false, then it would be the sort of thing that one could sensibly claim to know; but it isn't that sort of thing; so it isn't descriptive, but expressive.

The problem with this sort of defense is simply that it tries to defend one implausible claim by appealing to another; for why should we accept the claim that it is *meaningless* to claim to know that one is in pain? Is "I know I am in pain" meaningless in the way that "blah blah blah" is meaningless? No – each word in it is at least individually meaningful. Is it meaningless in the way "Truck blue Fred the" is? No, since unlike this sentence, it is perfectly grammatical (in the everyday sense of

"grammatical," that is, not Wittgenstein's technical sense). Nor does the sentence contain any explicit or implicit contradiction – another common basis for an accusation of meaninglessness. Well, is it meaningless in the way Chomsky's famous "Colorless green ideas sleep furiously" is? Perhaps this is closer to the sense of "meaningless" Wittgenstein has in mind, for Chomsky's sentence is full of what Ryle called "category mistakes," and it appears that Wittgenstein's idea is that to claim to know that one is in pain is to commit a category mistake. It is the sheer oddness of the Chomskian sentence that leads us to suspect that something or other is wrong with it, and indeed, the oddness of claiming either to know or to doubt that one is in pain is what Wittgenstein appeals to to support his claim.

But surely "I know I am in pain" and "I doubt I am in pain," though admittedly odd, are nowhere near the oddity of Chomsky's example. The oddity doesn't seem *semantic* (as it does in the Chomsky case) so much as *epistemic*. An apparently sincere (i.e. non-facetious) utterance of Chomsky's sentence leads us to suspect confusion or even madness on the part of the speaker; sincere utterances of "I know I am in pain," etc. sound merely pointless, too obvious or trivially true to be worth saying. "I know I'm in pain," though odd, isn't much odder than "I know that's my wife" – in most cases, a silly thing to say, though perfectly meaningful. In any case, where a statement does not appear to be meaningless by any of the ordinary criteria of

meaninglessness, the burden of proof is surely on the one claiming it is (literally) meaningless to show that it is; and I don't see that Wittgenstein has met that burden.[1]

My conclusion, then, is that Wittgenstein has, ultimately, no way to justify the claim that his commitment to premise (3) does not amount to an endorsement of logical behaviorism. And logical behaviorism is simply an unacceptable theory: It is, for one thing, subject to the same objection we saw other standard materialist theories were subject to, namely that it is logically possible that a system could fulfill the criteria it lays down for conscious experience – in this case, the instantiation of certain behavior or dispositions to behavior – and yet lack any conscious experience. (Wittgenstein would, no doubt, reject this sort of objection; but it is hard to see how he could do so in a non-question-begging way.) So, since logical behaviorism is false, and (3) would commit us to assuming that it is true, it follows that we have good grounds to reject (3), and thus to conclude that Wittgenstein's argument fails.

There is another important problem with logical behaviorism, though, and it is one which suggests another, final set of objections to Wittgenstein's position: Logical behaviorism notoriously fails to do justice to the *causal relations* holding between mental states and behavior. On the common sense view, my sensation of pain causes the pain behavior associated with it; which obviously implies that it is *distinct* from the pain behavior, contra logical behaviorism.

---

[1] We will have reason in the next chapter to look at Michael Lockwood's view that qualia can exist unsensed by, and apart from, any subject, and though I will argue against it, the weaker, related position that a subject can have qualia of which he is not aware does, as we shall see, have some

One commentator on Wittgenstein, Malcolm Budd, has recognized the importance of the causal role played by sensations such as pain, acknowledging that "it appears to be built into the language-game that the relation between a pain and its behavioral manifestation *is* causal" (1989, p. 70) and that this fact "places a constraint on the correct account of the concept of pain" (1989, p. 71). Given that Wittgenstein has ruled out the idea that sensations are private objects or qualia, it follows, Budd says, that the only way Wittgenstein could accommodate this fact would be to adopt a token-token mind-brain identity theory, on which sensations such as pains are causes of behavior (i.e. not identical to the behavior), though causes which are unobserved (but in principle observable) by the subject – the last element being important given Wittgenstein's insistence that first-person ascriptions of pain are not based on observation (since if they were, it would make sense to speak of *knowing* that one was in pain) (1989, pp. 71-72).

Still, Budd says, Wittgenstein would not have accepted this, for two reasons. First of all, any causal mechanism might in principle malfunction, in which case one might legitimately wonder whether or not the causal mechanism leading from bodily damage to pain behavior has failed to function properly, and thus might wonder whether he is in fact in pain even though he has no inclination to say that he is; for perhaps the pain has, due to some malfunction, failed to cause a self-ascription of pain. And this (perfectly intelligible) wondering would be incompatible with Wittgenstein's

---

plausibility. And if that view can be defended, it would obviously give good reason to suppose contra

insistence on the illegitimacy of either doubting or claiming to know one is in pain (1989, p. 73).[2]

Secondly, in line with his motivations for accepting premise (3), Wittgenstein would deny that the criteria for meaningful discourse about sensations such as pain could have anything to do with anything going on inside a person's body (which is, after all, generally unobserved by language users, even if observable in principle), any more than it could have anything to do with the goings-on of qualia in some immaterial, Cartesian theatre: Only manifest behavior and environmental circumstances play a role, in his view, in determining the meaning of sensation terms (Budd, 1989, p. 74-75).

The upshot of all this, Budd concludes, is that in implicitly rejecting the causal role of pains (and other sensations) as essential to the concept of pain, Wittgenstein's behavior-oriented (and ultimately behaviorist, as Budd more or less acknowledges) analysis amounts to "a revisionist account of the language-game played with names of sensations" (1989, p. 76). And while it isn't clear that Budd himself does so, I believe

---

Wittgenstein that one *could* sensibly doubt or claim to know that he had a particular sensation.

[2] Another way the mechanism could malfunction would be by causing an utterance of "I am in pain" when there is, in fact, no pain present. But Budd denies that *this* possibility would contradict Wittgenstein's claim that one cannot sensibly doubt that he is in pain, for, he suggests, "a sincere utterance of 'I am in pain' counts as a *self-ascription* of pain... only if it is the effect of the subject's [genuinely] *being in pain* [so that] the fact that a sincere utterance of the sentence could issue from something other than pain does not open the possibility that someone who understands the word 'pain' might falsely believe he is in pain" (1989, p. 74, emphasis mine). This suggestion is, to say the least, controversial: My utterance of "John is wearing a fez" on seeing him across the room is surely a genuine ascription to John of the property of fez-wearing, even if it is in fact caused, not by a fez being on his head, but rather by the upside-down trash can someone has placed on his head. Why should the case of pain be any different? But I needn't push this point here to make the case against Wittgenstein.

that this alone counts as a devastating objection to Wittgenstein's position, given his own conception of philosophical methodology, according to which "philosophy may in no way interfere with the actual use of language; it can in the end only describe it... It leaves everything as it is" (1953, sec. 124). The driving force of Wittgenstein's later philosophy is the idea that ordinary language is fine the way it is, and that philosophical problems arise only when we fail to understand how it actually works. So how can he possibly justify revising ordinary language – the very thing he criticizes traditional philosophy for doing – consistent with his own methodological scruples?

That it leads to this problem, and that it solidifies the case for interpreting Wittgenstein as a behaviorist, are damning enough consequences of the fact that pain bears a causal relation to pain behavior. But we can also appeal to this fact to make a final, direct challenge to premise (3). If it can be *part of the (ordinary language) concept* of pain that it is something which is caused by bodily damage and causes pain behavior, *even if it is something unobserved*, so that its not being (and generally *never* being) observed does not keep us from having meaningful discourse about it in a public language, what reason is there to rule out the possibility that *qualia* can be meaningfully discussed in a public language – indeed, to rule out, as Wittgenstein does, the idea that pain just *is* a quale? If we are to take seriously our *ordinary, everyday* sensation concepts, as Wittgenstein says we should, and those concepts allow for meaningful discourse about something never in fact observed, then they

allow for meaningful discourse about qualia.[3] The facts about ordinary usage thus not only fail to support Wittgenstein's case, they positively undermine it. Premise (3) of (my reconstruction of) Wittgenstein's private language argument against the existence of qualia thus collapses – and takes the entire argument along with it.

(b) Dennett's eliminativism

Daniel Dennett wants no more than Wittgenstein does to deny the existence of pains, itches, tickles, or sensory experiences in general; his eliminativism is thus in one respect not as radical as that of Churchland, who suggests that beliefs, desires and the like might be fictions on a par with witches and phlogiston (1981).[4] Like Wittgenstein, his aim is rather to show that there are no *qualia* involved in sensory experiences, no features accessible only from the subjective or first-person point of view, the reason being that the very concept of a quale is, he claims, a nonsensical one. Still, unlike Wittgenstein, he doesn't claim merely to be extricating common sense from a morass philosophical speculation has put it in; he takes the target of his attack to be a "pre-theoretical" and "intuitive" concept (1993, pp. 382-383) and implies that even Wittgenstein didn't go far enough given his allowance that a sensation is, though "not a *something*," nevertheless "not a *nothing* either" (1993, pp. 386-387).

---

[3] It might be claimed that the fact that qualia are not only never in fact observed, but are unobservable in principle, makes a crucial difference here. But it doesn't. Wittgenstein's case is based on the idea that only what *in fact* serve as publicly accessible criteria can play a role in determining the meaning of sensation terms – which is why *he himself* rules out (in principle observable) neural processes as well as (in principle unobservable) qualia as relevant. So far as Wittgenstein's own arguments are concerned, if they fail to rule out the former as relevant, they fail also to rule out the latter.

[4] Though interestingly enough, when it comes to qualia, as opposed to propositional attitudes, Churchland is less an eliminativist than a reductionist. See his 1989a.

Dennett's position (in his 1991 and 1993) comprises arguments for (i) a rejection of the notion of qualia itself, (ii) a rejection of some of the *attributes* commonly ascribed to qualia, particularly intrinsicality, and (iii) an account of perception which characterizes perceptual states as functional states whereby an organism discriminates objects in its environment. I am actually very much in sympathy with (ii) (at least as regards intrinsicality) and (iii), as will be very clear by the time chapter 6 rolls around, but though I think what he has to say under the rubric of (iii) is valuable (though I wouldn't, as he does, take what he says as a *replacement* for the concept of qualia), his defense of (ii) is, I think, (disappointingly) extremely weak, and his case for (i) is entirely unconvincing. We'll look at the problems with his arguments for (i) and (ii) now, saving for chapter 6 discussion of what is defensible in (ii) and (iii).

Dennett's first line of argument attempts to show that no sense can be made of the notion of *inverted* qualia, and that therefore no sense can be made of the notion of qualia simpliciter. Though he isn't entirely explicit about this, the argument appears to be one about identity conditions, the idea being that if we cannot meaningfully say that qualia are different, neither can we meaningfully say that they are identical, and thus we cannot meaningfully assert their existence at all. The argument also takes for granted a kind of *verificationism*, roughly the view that if there can be no evidence in principle for the existence of x, then x's existence cannot meaningfully be affirmed (Dennett, 1991, pp. 390, 461-462; 1993, p. 389). Now both verificationism and the

idea that significant assertion of the existence of something presupposes a statement of its identity conditions are controversial theses – certainly more controversial than Dennett himself lets on – but I hope to show that even if they are granted, Dennett's arguments fail.

Consider, as an example of a case of inverted qualia, the inverted spectrum hypothesis, according to which though our respective physiologies and behavior, verbal and otherwise, may be identical in all respects relevant to color, it is conceivable that what you see when you look at objects we both call red is what I see when I look at objects we both call green, and vice versa. Dennett argues that there is, in principle, no way of determining whether or not this scenario holds. For the sort of device we would need to construct in order to tell us whether or not someone's color qualia were inverted relative to our own (which he calls a "Brainstorm machine," after the movie *Brainstorm*) could be relied upon only if we could compare its deliverances with independently known facts about what other people's qualia were like, so as to make sure it was functioning properly.[5] But knowledge of what other people's qualia are like is just what the machine itself was supposed to provide us with in the first place. So construction of such a machine is impossible (Dennett, 1993, pp. 387-389; 1991, pp. 389-398). Our consequent inability to confirm or disconfirm the inverted qualia hypothesis, Dennett says (and here's where the verificationism comes in), shows "that

[5] See Meehl (1966) for an anticipation of the notion of such a device.

the very idea of inverted qualia is nonsense – and hence that the very idea of qualia is nonsense" (1991, p. 390).

Now the obvious rejoinder to this (again, granting verificationism for the sake of argument) is just to point out that despite my lack of evidence concerning what other people's qualia are like, I can still know what *my own* qualia are like. Furthermore, I can imagine that my color qualia at some point become inverted, so that what once looked red to me now looks green, etc. But since I would notice the difference, I would thereby have evidence of a qualia inversion.

But Dennett is well aware of this possible rejoinder – which he calls the *intra*personal inverted qualia scenario, to distinguish it from the more familiar *inter*personal scenario (1993, pp. 387-388) – and the bulk of his efforts are devoted to trying to undermine it. He uses two "intuition pumps," as he refers to his thought experiments, involving apparent inversions of qualia, to make his case. The first concerns a color qualia inversion of the sort just described, induced by neurosurgeons who tamper with your visual system (1993, pp. 387-388). He suggests that there are two ways in which they might produce the effect of getting you to think that your qualia have been inverted. They might do it by tampering with neural pathways that are early in the series involved in color vision, say in the optic nerve, so that your qualia are in fact inverted. But they might also do it by leaving those pathways (and thus your qualia) alone, and tampering instead with neural structures involved in memory, so that you misremember what your qualia were like before and mistakenly

believe your current qualia to differ from them. But then we are back in the same position we were in with the *inter*personal inverted qualia case, Dennett concludes. For your link to your past experiences via memory is, he says, just as hopeless an instrument for determining whether an inversion of your own past qualia has truly occurred as a "Brainstorm machine" would be in determining whether another's qualia have been inverted. So there is no way to confirm or disconfirm the hypothesis that your qualia have been inverted, and thus the *intra*personal inverted spectrum scenario, too, is nonsense.

At first glance, this argument seems clearly fallacious. For the way Dennett sets up the scenario not only allows for, but ensures that there are means of confirming whether or not your qualia were inverted: Just ask the neurosurgeons which sort of tampering they did, or, if you don't trust them, have someone else check your brain to find out. And since the scenario depends on the claim that there are (in principle at least) discoverable (if no doubt very general) correlations between certain states of the brain and memory (a plausible assumption in any case), so that the neurosurgeons can know just what tampering will do the trick, it cannot be the case that memory is relevantly like the Brainstorm machine. With the latter, there is supposed to be no way in principle to determine whether it is functioning correctly, and thus no way to determine what would count as genuine access to another person's experiences. With memory, Dennett's case depends on there being a way in principle to determine

whether it has been tampered with, and thus a way of determining what counts as genuine access to one's past experiences.

Still, there is a possible construal of the claim that memory is analogous to the Brainstorm machine that poses more of a challenge to the intrapersonal inverted spectrum scenario. It might be suggested that an appeal to neurophysiological evidence couldn't help one decide whether his qualia have been inverted *even given* the sort of correlations between states of the brain and memory that neuroscience can conceivably uncover. For such an appeal would presuppose the reliability of memory in that it would involve e.g. one's assuming that his memories of the neurophysiological evidence being of such-and-such a character are trustworthy. But then, just as in the case of the Brainstorm machine, one will have to assume just what he is trying to establish. That is, he will have to assume that his memory is reliable in order to establish that it reliably informs him of what his past qualia were like.

This version of the argument also fails, however. For this sort of problem about memory, unlike the problem with the Brainstorm machine, is not one that is peculiar to the inverted spectrum scenario. It is nothing but a special case of general skepticism about memory, and poses no problem for claims about qualia that it doesn't also pose for all our claims about the past. There is, then, a disanalogy between memory and the Brainstorm machine that undermines any attempt to argue that the intrapersonal inverted spectrum scenario suffers the same difficulty as the interpersonal scenario. The Brainstorm machine has *one* function, to inform us of what another's

qualia are like; so if there is no way in principle to determine whether it fulfills that one function, it cannot help us to determine whether another's qualia are inverted. But memory functions, not only to tell us about our past qualia, but about the past in general. So if we can justifiably believe that it functions reliably to tell us about the past *in general*, perhaps on the basis of neurophysiological evidence, etc. (and surely we think we can, unless we are skeptics about memory), then we can justifiably believe that it functions reliably to tell us about our past *qualia*, and thus that we can determine whether they have been inverted.[6]

So this thought experiment fails to show that there is no way of confirming or disconfirming the occurrence of a qualia inversion. But Dennett also develops another, on the surface more subtle, thought experiment to argue this. He asks us to consider Chase and Sanborn, two coffee tasters who work for Maxwell House (1993, pp. 389-399). After some years of working to ensure that the taste of Maxwell House coffee stays constant, and initially enjoying it, both men have come to dislike the taste of the coffee they have regularly to drink, but apparently for different reasons. Chase says that he no longer likes the taste of the coffee, because his tastes in coffee have changed. Sanborn says that his tastes in coffee haven't changed, but that he has come to dislike the taste of the coffee he samples because it now tastes different to him, a result, he thinks, of a change in his gustatory apparatus. Now both men are relying on memory in characterizing what has happened to them, and memory, of course, can be

[6] This construal of Dennett's argument thus founders on a claim involving memory which is similar

unreliable. So for each man, Dennett says, there are really three possible descriptions of what has happened, from the point of view of someone who believes there are qualia: (a) His qualia have stayed constant, and his reactive attitudes to them have changed; (b) His qualia have changed, and his reactive attitudes have stayed constant; (c) He is in some state intermediate between (a) and (b), that is, his qualia have changed somewhat and his reactive attitudes somewhat as well. Chase claims that (a) rightly describes his situation, but it may be that his memory is deceiving him and (b) or (c) is the more accurate description. Similarly, Sanborn claims that (b) is the correct description of his case, but perhaps (a) or (c) is what is really happening, unbeknownst to him.

The relevance of this scenario to the issue at hand can be seen by considering the question of how one might determine whether description (a), (b), or (c) is true of either man. Dennett grants that in the extreme cases of (a) and (b), there may be considerable evidence available to determine whether or not either of them holds of either man. Behavioral evidence, for example, might include Chase's ability or inability to reidentify various coffees, teas, wines, and the like in blind taste tests when only minutes intervene between sips, which could support or undermine his claim to know that his qualia have not changed. There might also be physiological evidence, such as anomalies or the lack thereof in Sanborn's gustatory apparatus. "But such indirect testing," Dennett says, "cannot be expected to resolve the issue when the

---

to the one we've seen Wittgenstein's argument appeals to.

effects are relatively small – when, for instance, our rival hypotheses are Chase's preferred hypothesis (a) and the minor variant to the effect that his qualia have shifted *a little* and his standards *less than he thinks*" (1993, p. 394). It is in such cases, he thinks, that difficulties lie for the claim that there are means of confirming or disconfirming the occurrence of a qualia inversion.

To convince us of such difficulties, Dennett asks us to imagine, further, a case in which Chase's qualia are inverted due to neurosurgery, so that sugar tastes salty to him, salt tastes sweet, and so forth, but in which he nonetheless eventually adapts to the inversion, so that his behavior becomes indistinguishable from what it was before. Even if the physiological evidence tells us that the adaptation is due to changes in his memory accessing process, Dennett says, there are still two possible construals of these changes: (I) Chase's current qualia are still inverted relative to his old ones, but the changes in his memory accessing process have altered his memories of the latter, so that he now fails to notice the difference; and (II) "The memory-comparison step occurs just prior to the qualia phase in taste perception; thanks to the revision, it now *yields* the same old qualia for the same stimulation" (1993, p. 395). Neither introspection on Chase's part nor physiological evidence can help us decide between (I) and (II), according to Dennett. Therefore, the claim that Chase's qualia are still inverted, and thus the claim that he has qualia at all, cannot be either confirmed or disconfirmed, and is thus meaningless or nonsensical (again, granted verificationism).

Now it is not, I think, entirely clear what interpretation (II) is supposed to amount to. If we're talking about memory comparison, it is hard to see how this could occur "prior to the qualia phase," since if no qualia have yet been produced, what is there to compare to the memory of earlier qualia? At any rate, we are to suppose that the memory process somehow interferes with the qualia generating process so that the qualia actually produced are similar to those Chase had before the inversion. Presumably, then, what the memory process does on interpretation (II) is just reinvert the qualia. But given what Dennett has already granted, namely that cases like (a) and (b) are arguably capable in principle of neurophysiological confirmation or disconfirmation, why suppose the same can't be said of (I) and (II), which appear to differ (in confirmability or disconfirmability) from (a) and (b) only in degree?

The idea appears to be that there might be neural structures underlying memory which would affect a subject's judgements about qualia, and yet the operation of which would be equally well interpretable in terms either of (I) or (II) – that is, that the description of the relevant neural structures neuroscience gives us would not be fine-grained enough to allow us to favor one description over the other. This seems a fairly harmless supposition, but it is puzzling why Dennett should think that it, all by itself, poses any difficulty for the advocate of qualia. First of all, even if there *might* be neural structures whose neuroscientific description is indeterminate in the way described, Dennett's given us no reason to think there *are* any. But more to the point, even if there are, what Dennett would need for a verificationist argument to go

through is not merely a neuroscientific description which is *in fact* indeterminate – after all, this might be merely a reflection of current neuroscientific ignorance – nor even one which *in practice*, given economic or technological constraints, could not be made more determinate, but one which *in principle must be* indeterminate. And what reason do we have for thinking it even *likely* that we'll discover a structure which can be given only this sort of indeterminate neuroscientific description, much less that there *must* be one? As Owen Flanagan shows (1992, pp. 78-79), we can certainly *imagine* circumstances, in the case of (I) and (II), in which we would be led by empirical considerations to accept one over the other. It might turn out, for instance, that Chase's pre-inversion reports of saltiness and sweetness correlated with two particular neuronal firing patterns, and that these firing patterns were inverted relative to his immediate post-inversion reports, so that there is reason to believe that these firing patterns were responsible for salty taste qualia and sweet taste qualia respectively. Given that memory areas in the brain are responsible for the later adjustment, then if these neuronal firing patterns remain inverted, we will have reason to believe that memory is having its effect late in the perceptual process, and that (I) is correct; while if the firing patterns are no longer inverted, this will favor the hypothesis that memory is working early in the process, and thus (II). What rules such a possibility out?

It appears, then, that both of Dennett's "intuition pumps" fall far short of showing "that the very idea of inverted qualia is nonsense – and hence that the idea of

qualia is nonsense" (1991, p. 390), even if he is granted their (controversial) verificationist presuppositions. But he also appeals to them to fulfill his second, less ambitious, aim of showing that some of the properties traditionally ascribed to qualia are questionable, even if the existence of qualia (of some sort) is defensible. Of the victim of the inverted-spectrum-producing surgery, he writes:

> [N]othing in the subject's experience can favour one of the hypotheses over the other. So unless he seeks outside help, the state of his own qualia must be as unknowable to him as the state of anyone else's qualia; hardly the privileged access or immediate acquaintance or direct apprehension the friends of qualia had supposed 'phenomenal features' to enjoy!...[W]e cannot tell in our own cases whether our qualia have been inverted – at least not by introspection. (1993, p. 389)[7]

So Dennett takes his thought experiments to show at least (1) that one cannot determine (conclusively) by introspection whether his qualia have been inverted (even if he may do so by other, "outside" means), and thus (2) that qualia are not, after all, directly apprehensible in consciousness, as has traditionally been thought.

Unfortunately for Dennett, the intuition pumps under consideration seem scarcely more impressive when we read them as supporting only (1) and (2). Let us grant that they do support (1) – though only with a heavy emphasis on the qualifier "conclusively," since for all Dennett has shown, detecting a qualia inversion by

[7] Dennett also draws a similar moral from the Chase and Sanborn thought experiment (1993, p. 396).

introspection is no more problematic than any other epistemic practice involving memory. Does (2) follow from (1)? It isn't obvious how it's supposed to. To claim that qualia are directly apprehensible is not to claim that our memories are infallible with respect to them. Yet the falsity of the latter is the most that Dennett can claim to have shown. Nor is this a particularly interesting result: Who ever said memory was infallible in *any* respect? To claim that qualia are directly apprehensible is just to claim that I know directly the qualia that I am currently having. I may not know directly, or at all, whether they are the same as or different from qualia I have had in the past.[8]

But perhaps Dennett's argument could be beefed up as follows: Traditionally, direct apprehensibility was supposed to yield *incorrigibility*, in that it was supposed to ensure that one's judgements about his own qualia cannot be in error. The idea would be that *direct* apprehensibility should, if qualia have it, entail immunity to error by removing the sources of error associated with *indirect* access to an object of knowledge: Maybe mirrors, clouds, and the like could distort my perceptions in the case of the apprehension of physical objects, but there should be nothing analogous in the case of the inner eye's apprehension of qualia if that apprehension is direct. But since it is conceivable that I might be misremembering my past qualia, e.g. in thinking they were inverted when they were not, or vice versa, it follows that my judgements about my current qualia as regards their similarity or dissimilarity to past qualia might

[8] See Block, 1994, p. 212, for a similar criticism of another reconstruction of Dennett's argument.

be mistaken, so that they are not incorrigible after all. And therefore my qualia cannot be directly apprehensible after all.

Even this (rather charitable) reconstruction of Dennett's argument won't do, however. Direct apprehensibility *doesn't* entail incorrigibility *without qualification*; it entails, at most, immunity to error *of the sort* that results from apprehending an object of knowledge only indirectly. And Dennett's thought experiments don't show that qualia are not incorrigible in *that* sense; the problems they pose for our judgements about qualia have rather to do with memory. In the case of one's judgement that what he sees before him is a tree, even if his memory of what a tree looks like is known to be correct, there are still many potential sources of error deriving from his merely indirect, at best, apprehension of the tree.[9] But for all Dennett's thought experiments show, given that one's memory can be trusted, his judgements about his own present qualia are otherwise incorrigible. So he's given no reason to doubt that qualia are directly apprehensible.

It might also be conjectured that in inferring (2) from (1), Dennett may be arguing along lines similar to those of Wittgenstein's private language argument. That is, he may be arguing that one's ability to *identify* a current quale depends on his ability to compare it to an earlier quale, so that if he cannot know through

[9] This point doesn't depend on an *indirect realist* theory of perception (though I will in fact be defending such a theory in chapter 4). The point is just that in the case of perception of external objects, it is always possible that something might quite literally get between the observer and the object of perception, thereby producing errors in one's judgements about that object – a fact no direct realist would deny; while nothing analogous should be possible in the case of introspection of qualia if qualia are directly apprehensible.

introspection whether or not he can trust his memories of what his earlier qualia were like, he cannot identify, through introspection alone, a current quale, and so he cannot be said directly to apprehend the latter. This won't work either, though, even apart from all the objections which, as we've already seen, the Wittgensteinian argument is subject to. For even if, given the possibility that one's perceptual apparatus and memory have been tampered with, one cannot know beyond the possibility of doubt that the color quale he is currently having is what he previously called "blue" without performing third-person checks, one can still know that he is having *some* quale or other through introspection alone. That is, he does not have to infer the *existence* of the quale, but can apprehend it directly, even if inference (on the basis of third-person, neurophysiological evidence) would be required for one to *identify* the quale as of the same sort one had previously encountered.

The failure of Dennett's arguments against direct apprehensibility is as significant as the failure of his arguments against qualia themselves, as far as his attempt to *dissolve* the qualia problem is concerned. For as I've characterized that problem, it is precisely a problem about there being features of the mind which are accessible only from the subjective or first-person point of view, and which cannot be captured within the third-person descriptions of physical science as typically understood. And the idea of directly apprehensible features of the mind just is the idea of features which can only be accessed from the first-person point of view, which,

unlike third-person phenomena, need not be accessed only through the mediation of sensory organs, photons, sound waves, and the like.

What Dennett seems most concerned to refute, however, is the idea that qualia, even if they exist, could be anything *over and above* a perceiving subject's reactive dispositions and discriminative capacities, that is, that they could be *intrinsic* properties. But unfortunately, Dennett's arguments against intrinsicality are, in my view, the weakest of all his arguments. I say "unfortunately" because, again, this is the part of Dennett's position with which I am most sympathetic; and in chapter 6 I will try to defend the claim that qualia are not intrinsic properties – though it will turn out that the version of this claim I will defend is radically different from Dennett's. Suffice it for now to say that, in any case, the denial of intrinsicality *by itself* does not either solve or dissolve the qualia problem: Even if qualia are not intrinsic properties, they might nevertheless be directly accessible only from the first-person point of view. Exactly how an affirmation of the subjectivity and direct apprehensibility of qualia might fit together with a denial of their intrinsicality is an issue that will have to wait for chapter 6, though.

For now, let us examine Dennett's arguments against intrinsicality. He begins by again appealing to a couple of intuition pumps: The first involves a beer drinker who hates the taste of beer when he first tries it, but eventually learns to like it (Dennett, 1993, p. 397; 1991, pp. 395-396). The second involves a hater of cauliflower who takes a pill which causes him to come to enjoy eating it (1993, p.

399). The first shows, he claims, that "prolonged beer drinking leads people to experience a taste they enjoy, but precisely their enjoying the taste guarantees it is not the taste they first experience," which means that the taste of beer is constituted by one's dispositions to react to it, and thus is not an intrinsic quality (1993, p. 397). The second also shows that tastes are not intrinsic qualities, because the taster "will find nothing *in [his] experience* to shed light on the question" of whether the taste of the cauliflower has stayed the same, and he has now come to like it, or whether it has changed (1993, p. 399); the idea – apparently – being the verificationist one that if the taste of the cauliflower were an intrinsic quality, something more than one's dispositions to react to it, then there should be some way to verify the claim that the taste of the cauliflower pre-pill and post-pill is the same, despite the change in reactive dispositions.

I must confess that I find these arguments simply baffling. For there seems just no (non-question-begging) reason at all to accept the claims that, in the case of beer drinkers, "precisely their enjoying the taste [of beer] guarantees it is not the taste they first experience," and in the case of the cauliflower eater, "there is nothing in his experience to shed light on the question" of whether the taste is the same as before. Indeed, there is an obvious reason *not* to accept them – one provided by Dennett himself when he says, of the cauliflower eater, "of course [he] recognize[s] that the taste is (sort of) the same – the pill has not made the cauliflower taste like chocolate cake, after all..." (1993, p. 399). "Of course" indeed! Who, apart perhaps from a

character in one of Dennett's "intuition pumps," is going to say that the taste of his first beer or first plate of cauliflower and that of his last is different? Even if it were granted that it is *somewhat* different, that wouldn't be enough for Dennett's argument. As long as it is "(sort of) the same," as long as someone would clearly notice *something* common to the first and last taste experiences in the beer or cauliflower cases, there *would* be grounds (and grounds *in the subject's experience*) for saying that the later, more positive reactions are reactions to *the same taste.* So not only does Dennett here put forward arguments with controversial, unsupported premises; he then makes concessions which undermine any plausible defense of those premises!

Dennett also appeals to the different reactions people have to the chemical substance phenol-thio-urea to argue against the notion of intrinsicality (1993, pp. 397-398; 1991, p. 379). To about three quarters of the human race, this substance tastes bitter, and to the rest it is tasteless. If, through selective breeding, we weeded out the latter, phenol-thio-urea would be regarded as paradigmatically bitter; and if we weeded out the former, it would be regarded as paradigmatically tasteless. So, Dennett says, whether it is bitter or tasteless is relative to the perceptual equipment of those who taste it. "Clearly, public bitterness or tastelessness is not an intrinsic property of phenol-thio-urea but a relational property, since the property is changed by a change in the reference class of normal detectors" (1993, p. 398).

It is not entirely clear what Dennett takes this example to show, given, for instance, that it is "*public* bitterness or tastelessness" that is shown not to be intrinsic.

Is it supposed to show that there are no intrinsic qualities *whatsoever* involved when one tastes phenol-thio-urea? If so, it fails, since this claim doesn't follow from the facts described at all. All they show is that whether the substance in question causes one to have this or that quale depends on the character of one's perceptual apparatus. And this is consistent with the claim that, given that one's perceptual equipment is such that he is caused to have, say, bitter qualia when tasting it, those qualia have the character they do intrinsically, independently of the subject's reactive dispositions and the like. Is it supposed to show, instead, only that "[p]roperties that 'seem intrinsic' at first often turn out on more careful analysis to be relational" (1993, p. 397), in this case that the "public" property of causing people to have certain experiences, which might seem to be a property intrinsic to a chemical substance, is in fact a relational property? In this case, the argument would seem to be: The (admittedly public) property of *causing* certain experiences was thought to be intrinsic, but is in fact relational; so perhaps (allegedly private) properties, like the quale or *character* of a certain experience, though they seem intrinsic, are also relational. Well, maybe so; no one ever said that all properties thought to be intrinsic must really be so. But, then again, maybe not. By itself, such an argument shows at most only that qualia *may* turn out not to be intrinsic. It provides no reason at all for thinking that in fact they are not.[10]

---

[10] Dennett also suggests that experiments involving the inversion of a subject's visual field by means of special goggles show that the "right-side-upness" of one's visual field is a relational, not an intrinsic, property of one's visual field, and that this supports his claim that qualia are not intrinsic properties either (1993, pp. 399-400; 1991, pp. 397-398). But how the former supports the latter is as

Perhaps underlying Dennett's arguments against intrinsicality is the thought that since, as we've seen, there may be cases where evidence of a neurophysiological or behavioral sort is required in order to decide between hypotheses concerning qualia, having a certain quale must involve being in a certain brain state or being disposed to certain kinds of behavior. And from this, it might be thought, it follows that qualia are at least partly *constituted* by brain states and/or behavior, and so are not intrinsic. But this doesn't follow: The advocate of intrinsicality need, in the face of just *these* facts, grant only that qualia are *causally correlated* with brain states and behavior, not that they are *constituted* by them. Even if a quale is so correlated with a certain kind of brain state or behavioral disposition, it might still be something *over and above* them, and thus an intrinsic property.

These particular arguments are, in my view, bad enough that it would be nice to have an *explanation* for why Dennett should think them worth putting forward. My suspicion is that part of Dennett's aim is at least to produce intuitions in his readers (i.e. hence the use of what he calls "intuition pumps") that run counter to the common sense ones which give rise to the notion of qualia. Perhaps his thought experiments are intended, not decisively to refute his opponent, but only to rattle him, to shake his confidence in the intuitions underlying the belief in intrinsic properties. But given the objections we've seen his thought experiments are open to, it's not clear even this more limited goal has been fulfilled. Alternatively, Dennett may intend that these

---

unclear as how the phenol-thio-urea case is supposed to undermine intrinsicality. (And see Kirk

arguments be read in light of his commitment to what he in other contexts calls "the method of heterophenomenology" (1991, p. 72), according to which any acceptable theory of the mind "will have to be constructed from the third-person point of view" (1991, p. 71). Indeed, the beer drinker and cauliflower arguments are much more plausible granted this methodological scruple: Given *just* third-person evidence, it *may* indeed be difficult if not impossible to tell whether there is anything to the taste of something over and above a person's reactive dispositions to it. The problem, of course, is that Dennett's arguments thus become even more blatantly question-begging against the advocate of qualia, whose position is precisely that there is more to the mind than third-person phenomena.

Dennett has one final argument against the existence of qualia which draws on all the thought experiments we've looked at so far, which, he alleges, together show that our ordinary, pre-theoretical ways of talking about experiences, upon which the notion of qualia rests, are "confused and potentially incoherent" (1993, p. 398). In particular, he says, they show that the best we can do with our everyday concepts is distinguish between our past experiences and our current ones. But the moment we try to determine in what respect they are different, that is, whether it is the character of our experiences or our reactive dispositions toward them which have changed, we run into trouble. In regard to the cauliflower case, Dennett says that the suggestion that the cauliflower will taste just the way it always did, and yet the eater will now like that

---

1994, pp. 55-66 for discussion of some other problems with Dennett's use of this example.)

taste, may be self-contradictory (since, on Dennett's view, if he now likes the taste, it just can't be the *same* taste as that encountered before) (1993, p. 399); and in regard to the Chase and Sanborn case, he says that "[w]hen Chase thinks of 'that taste' he thinks equivocally or vaguely" (1993, p. 398). Thus, for instance, Chase, when apprised of the possibilities, is led in two directions: He doesn't know whether to say that the coffee tastes the same as it did (and his reactions have changed), or that it doesn't (and his reactions have stayed constant). "His state then and his state now are different – *that* he can avow with confidence – but he has no 'immediate' resources for making a finer distinction..." (1993, pp. 398-399). Dennett's argument here, as far as I can tell, is that since we say things like "The cauliflower tastes the way it did when I hated it, but now I like it," which he says is self-contradictory, and since Chase is led to say both that the coffee tastes the same and that it does not, and both claims can't be true, there is some incoherence in the very idea that there are properties, e.g. tastes, understood as qualia, that exist over and above one's reactive dispositions, and persist when they change. The ordinary notions that underlie the concept of qualia, he claims, here lead us into contradictions.

This argument appears to rest on an equivocation and on mistaking indecisiveness for self-contradiction. Dennett says that common sense provides us with no way in principle of deciding whether or not the coffee tastes the same, that Chase "has no 'immediate' resources for making a finer distinction," and is thus led into incoherence, affirming both that it does taste the same and that it does not. But in

doing so, he overlooks the fact that there is a perfectly good distinction in ordinary language between senses of "taste," and that when these are kept in mind, no incoherence such as the one Dennett says Chase must fall into need result. That is, there is a sense of "taste" in which something is said to taste either good or bad, and a sense of "taste" in which things are said to taste like chicken, or like chocolate, or whatever. That this is a real, everyday distinction is clear from the case of water: In the latter sense, we say that it has no taste, while in the former sense, we often say that it tastes good on a hot day. Consider also that when someone says that chicken and chocolate both "taste good," he is not thereby saying that they taste the same, and thus wouldn't be contradicting himself by going on to assert that chicken and chocolate taste very different. For again, in asserting that things "taste good," we are using "taste" in a sense different from that it has when we say something "tastes like" something or other. In Chase's case, what is obviously happening is not that he is led to affirm two contradictory claims, "It tastes the same" and "It doesn't taste the same," which would be incoherent. Rather, he is (at first) led to say "It tastes the same in the second sense (i.e. like the coffee Chase tasted years ago), but not in the first sense (i.e. now it tastes bad)," which is perfectly coherent. The same goes for the cauliflower case: That "It tastes the same as it did, but I now like that taste" is not self-contradictory is clear when we keep in mind that there are two senses of "taste" involved. Nor can Chase be said to be contradicting himself later, when, on considering that his sensory apparatus may have been altered, he comes to wonder

whether the coffee does after all taste the same (in the second sense). He is not, in this case, led to affirm two incompatible statements, but just reluctant to draw a firm conclusion when made to consider the possibility that his memory is failing him. And if he wants to draw a conclusion, there is in fact a way for him to acquire evidence that would enable him to do so, for we have good reason to believe that, in principle, neuroscience can provide a means of deciding between the possible descriptions of Chase's condition. At any rate, as we have already seen, Dennett has failed to show otherwise.

Now perhaps Dennett's aim here isn't what I've been taking it to be. For though he does say things that lead one to believe that he is claiming that Chase and the other characters in his thought experiments are contradicting themselves, he also says things that may tell against this interpretation. He says, for instance, that Chase speaks "equivocally or vaguely," as we have seen, and if Chase's statements are equivocal or vague, then it seems that they are thereby less clearly contradictory. So perhaps Dennett's claim is not that our everyday concepts regarding our experiences lead us into outright contradiction (despite his explicit talk of contradiction and incoherence), but rather just that those concepts are loose or ill-defined. If so, this would make for a much less ambitious argument than the one I have been considering. To show that a set of concepts leads us into contradiction may indeed be to provide a decisive reason to abandon those concepts. But to show a set of concepts to be merely ill-defined is not at all necessarily to show that there is anything wrong with

them: Despite the famous paradox that results from the vaguenes of the concept of a heap, no one doubts that there are such things as heaps, and that the concept of a heap has clear applications. At any rate, if Dennett *is* merely trying to show the concepts underlying the notion of qualia to be ill-defined, he has failed to do even that, for even read in this way, his argument still has the flaws noted above.

I conclude from this examination of Wittgenstein's and Dennett's arguments – fairly representative of attempts to *dissolve* the qualia problem, as opposed to solving it – that such attempts have no force at all. Ironically, it will nevertheless turn out that a Dennett-like positive account of the nature of perceptual states plays an important role in *solving* the qualia problem. But before seeing how, we must first examine some failed attempts to do so.

3. Attempts to solve the qualia problem

(a) Knowing qualia: the direct introspection of brain states

If qualia are, contra Wittgenstein and Dennett, real, and physicalism is true, then facts about qualia must really be, despite appearances, physical facts. Indeed, Paul Churchland insists, they are nothing but facts about the brain. He argues, along more old-fashioned materialist lines – without appealing to functional role and the like – that qualia *just are* brain states: Mary's introspection of her red qualia on leaving the black and white room is nothing but the *direct introspection of states of her brain* (1989a, 1989b). But how can this be?

Churchland's answer (an answer which supplements his other reply to Jackson, which we considered earlier) is, following Robert Van Gulick's (1993) classification, representative of one of three sorts of response to the knowledge argument given by physicalists who grant that Mary gains knowledge when she leaves the room: the response that this only seems impossible, and Jackson's argument only seems sound, because of an equivocation on "knows" which anti-physicalists commit when they contrast knowing facts about qualia with knowing neurophysiological facts (and which occurs between premises (1) and (2) of the knowledge argument). In the terms of Russell's (1988, pp. 46-59) classic distinction: what Mary has when she's still in the room is knowledge by *description*, and what she gains on leaving it is knowledge by *acquaintance*. But having only the one kind of knowledge at first and later gaining the other is consistent with *one and the same thing* being known in both cases.

Churchland further glosses the notion of knowledge by acquaintance in good materialist fashion by describing it as "a matter of having a representation of [in Mary's case] redness in some prelinguistic or sublinguisitic medium of representation for sensory variables, or to be a matter of being able to *make* certain sensory discriminations, or something along these lines" (1989a, p. 62).[1] Indeed, all having a quale amounts to in Churchland's view is the having of such a representation (cashed out in terms of having a certain set of neural connections) or discriminative ability. And since Mary knows everything physical there is to know about such things before leaving the room, she already knows by description all about qualia. What she has knowledge of after leaving the room is not something she didn't know before; rather, she just knows, *in a new way* – by acquaintance – something she already knew before.

This reply is a little confusing in that it is not entirely clear from Churchland's account *exactly what* is supposed to be known by acquaintance when one knows a red quale, if this way of having knowledge by acquaintance *of a quale* is itself a matter of having a representation *of redness*, and the quale *itself* is supposed to be that representation. Is it an objective property of the red object itself (i.e. whatever property causes the neural state Churchland would identify with the quale) that one knows by acquaintance? Or is it the quale itself, that is, a state of the perceiver's brain? Presumably what he has in mind is that to have a red quale is just to have a

[1] Clearly, what Churchland means by "redness" here is just the *objective* feature or features possessed by a physical object by virtue of which it produces a *sensation* of red in a perceiving subject, a sensation Churchland would identify with a brain state or process, rather than with some subjective, first-person feature, as a critic of materialism would.

representation of a feature of a red object, understood as a certain kind of brain state, and to introspect that quale is just to have a higher-order brain state which itself serves as a representation of the lower-order brain state or quale. So perhaps knowledge through perception of a red object and knowledge of the brain state through which that perceptual knowledge is had are, on Churchland's view, both cases of knowledge by acquaintance.

In any case, Churchland's position has more serious problems than this ambiguity. Whatever intuitive force this reply to the knowledge argument has rests, I think, on the assumption that it appeals to a distinction, that between knowledge by acquaintance and knowledge by description, that is uncontroversial, neutral in its significance as between materialism and opposing views. But the falsity of this assumption can, I think, be made evident by considering a parallel reply to the zombie argument: Churchland might say that to imagine a world physically identical to ours would *just be* to imagine a world with qualia, since qualia just are brain states. And he might defend this reply as non-question-begging by saying that we know qualia, after all, by introspection, by direct *acquaintance* with them; and that this process is nothing but the having of a representation of a feature of a red object, understood as the having of a certain kind of neural state (or, as we've seen, perhaps the having of a representation of this first representation, the second-order representation itself being just a higher-order brain state). But of course, this defense would *itself* obviously be question-begging: the proponent of the zombie argument would say, rightly, that the

whole point of his argument is that *any* brain state, even one caused by certain features of physical objects and arguably serving as a representation of those features, could possibly exist apart from qualia.

The problem is that the way one cashes out the notion of knowledge by acquaintance itself depends on whether or not one takes qualia to be physical properties. Nor should this be surprising: Russell himself, when he made famous the distinction between knowledge by acquaintance and knowledge by description, characterized the former as knowledge of *sense-data*, of what would today more commonly be called qualia, understood, as they traditionally are, as subjective, first-person properties; indeed, on his view, sense-data or qualia are the *only* things we ever know by acquaintance, and never *physical* objects or processes themselves (Russell 1988, p. 47)![2] So it will hardly do for Churchland to appeal to this distinction as though it were neutral territory from which to launch a challenge to the knowledge argument. If knowledge by acquaintance *just is* knowledge of qualia, understood as irreducibly subjective, first-person features, then it's no refutation of Jackson to argue that all Mary gains knowledge of on leaving the room is knowledge by acquaintance; for in that case, there *would* be something over and above the physical facts, all of which she already knew while in the room. And if Churchland would deny that it is (as of course he would), then the burden of proof is on him first to give some non-

[2] It is true, of course, that Russell went on, later in his career, to identify sense-data or qualia with states of the brain; but as we'll see in chapter 4, his mind-brain identity theory was anything but a *physicalist* theory of the sort Churchland would sympathize with.

question-begging reason to assume that it isn't, before appealing to the notion in order to reply to Jackson.

As we'll see in the next chapter, there are very good grounds for thinking knowledge by acquaintance is precisely what Russell said it was. Part of Russell's case for the existence of sense-data or qualia, and for saying that they are all we ever know directly, by acquaintance, consisted of the sorts of considerations regarding perception that I alluded to when developing the notion of qualia in chapter 1, and will discuss at length later on. As will then become clear, and is already starting to come out here, the qualia problem is tightly bound up with the problem of perception (and the related problem of skepticism), so that it is impossible to do justice to the former without addressing also the latter. (Indeed, it is surprising, and unfortunate for their positions, that contemporary writers on the mind-body problem – unlike earlier philosophers from Descartes and Hume down to Russell – so often theorize about it without considering what bearing those other issues might have on it.)

There is, at any rate, also reason to doubt Churchland's supposition that Mary doesn't, on leaving the room, gain any knowledge that might be characterized as knowledge by description. For the latter sort of knowledge is, if nothing else, knowledge of propositions[3]; and there are grounds for thinking that Mary does indeed gain some such knowledge on leaving the room, even though she had (as Churchland grants, at least for the sake of argument, 1989a, p. 63) all the propositional knowledge

there was to have about the brain. The philosopher of language David Kaplan has written:

> Many of our beliefs have the form: "The color of her hair is _____," or "The song he was singing went _____," where the blanks are filled with images, sensory impressions, or what have you, but certainly not words. If we cannot even *say* it with words but have to paint it or sing it, we certainly cannot believe it with words. (1990, p. 387)

This suggests that there are *propositions* some of the constituents of which are images, sensory impressions, and the like (rather than merely Fregean senses, or Russellian objects, or whatever else one's favored theory of propositions would allow for). And if there are, surely the object of Mary's newly acquired knowledge upon leaving the room would seem precisely to be one of them, namely the one expressed by the sentence (which she thinks to herself) "A red apple looks like _____," where the blank is filled by a reddish quale. So it is arguable that Mary does gain new knowledge by description upon leaving the room, in which case, contra Churchland, there need be no equivocation on "knows" between premises (1) and (2) of the knowledge argument, and the argument goes through.

Fully to defend this suggestion would require an excursion far into the forbidding jungles of contemporary philosophy of language, and that is not an excursion I care to take just here and now. Suffice it to say that it at least adds to the

---

[3] This is (part of) what distinguishes it from knowledge by acquaintance, on at least some construals

case, already made on other grounds, for rejecting Churchland's solution to the problem of qualia.

(b) Knowing qualia: "knowing how," not "knowing that"

The second sort of reply to the knowledge argument suggested by physicalists who accept that Mary gains knowledge on leaving the room involves a different distinction between kinds of knowledge than that employed by Churchland, namely the distinction between "knowledge how" and "knowledge that," where the second sort of knowledge involves knowledge of facts, but the first does not, and involves instead the having of certain *abilities*. The idea here, as with Churchland's suggestion, is that the knowledge argument equivocates on "know" between premises (1) and (2). What Mary gains when she learns "what it's like" to see red is not knowledge of facts, even of facts she knew before in a different way, but rather certain abilities. As David Lewis, one proponent of this view, puts it:

> [K]nowing what it's like is not the possession of information at all. It isn't the elimination of any hitherto unknown possibilities. Rather, knowing what it's like is the possession of abilities: abilities to recognize, abilities to imagine, abilities to predict one's behavior by means of imaginative experiments. (Someone who knows what it's like to taste Vegemite can easily and reliably predict whether he would eat a second helping of Vegemite ice cream.) (1991, p. 234)

---

of the distinction. (Cf. Martens, 1992)

This suggestion has the advantage that it does not, as Churchland's view does, appeal to a distinction the proper characterization of which itself is a matter of dispute between physicalists and their opponents. But it has other problems, which, as in the case of Churchland's position, become clear when we try to apply it to a criticism of the zombie argument. To that argument, Lewis could perhaps respond by saying that, since knowledge of qualia is just the having, by a certain kind of physical system, of certain abilities, it is impossible that there be a world physically identical to ours and yet lacking qualia. But this would be a completely unconvincing objection, since it appears perfectly conceivable (and thus possible) that a system could have just the abilities we have, including certain cognitive abilities to predict future behavior, etc. (allowing at least for the sake of argument that such abilities themselves are explicable in purely functional terms), and yet lack any conscious experiences whatsoever.

In the scenario presented in Jackson's version of the knowledge argument, what Mary knows while still in the room (or what we could take her to know, slightly to expand on Jackson's example without altering its basic idea) would more or less be all the facts of the sort that would obtain in the zombie world, which would include facts about people's abilities, including her own, since those facts are just further physical facts: she'd be able to predict (as far as it would be possible to predict the behavior of any complex physical system) what objects people would discriminate and (given that there are interesting correlations between mental states and brain states) even which of their own mental states they'd be able to discriminate, and so forth; and

she could predict the same sorts of things about *herself*. And thus she *would* have the abilities Lewis talks about. But for all that, she would still learn something on her release, in learning what it's like to see red. Therefore, learning what it's like is *not* (or not merely) the gaining of certain abilities.

Now someone might respond by saying that nevertheless, Mary, while in the room, would lack *some* abilities which she gains on leaving, e.g. the ability to imagine what red looks like, etc., so that it is still an open possibility that what she gains on leaving the room *is* merely an ability. But the problem with this is that it seems clear that the abilities she would lack would just be abilities of the sort one can have only after gaining certain kinds of propositional knowledge or "knowledge that," in this case the knowledge *that red looks a certain way* (namely *this* way, the way it looks as she gazes at the apple for the first time). And if the physical facts are all the facts there are, she should have had this knowledge while in the room. (Recall the point I made earlier, inspired by Kaplan's suggestion, about Mary gaining new propositional knowledge, which would be "knowledge that," namely the knowledge that "A red apple looks like _____.") Again, learning what it's like involves more than the gaining of certain abilities, and it seems to involve even that only because it also involves the gaining of new factual knowledge.

All told, then, this attempt to undermine the knowledge argument by claiming it commits an equivocation on "knows" seems no more successful than Churchland's.

(c) Knowing qualia: knowing physical facts under new concepts

The last of the three sorts of reply to the knowledge argument distinguished by Van Gulick (1993) is the one he himself prefers. On this view, what Mary learns on leaving the room is a new concept. This new concept allows her to come to know new propositions, but only on a fine-grained scheme of individuating propositions. What this means is best explained by reference to examples: The proposition that 5 + 7 = 12 and the proposition that 38 is the square root of 1,444 are the same proposition on a *course-grained* scheme of individuating propositions (i.e. one that takes propositions to be functions from possible worlds to truth values); but they are different propositions on a *fine-grained* scheme, one that takes account of the differences in conceptual constituent structure between these two propositions. Another example would involve the propositions that water freezes at 32 degrees Fahrenheit and that H2O freezes at 32 degrees Fahrenheit, where they are the same proposition or different propositions depending on whether they're individuated on a coarse- or fine-grained mode of individuation. In this case, a fine-grained individuation would be one that took account of the difference between the concepts associated with "water" and "H2O," respectively. The idea is that just as *these* two concepts apply to the same thing (namely water), despite the fact that they are different concepts, so too might what the concept Mary learns on leaving the room apply to be the very same thing as that to which concepts she already had while in the room applied; and accordingly, the propositions she learns might be about exactly the

same facts she already knew about while in the room. And those facts might, for all Jackson has shown, be physical facts, so that the argument fails to refute physicalism after all.

What Van Gulick is suggesting here, then, is that the fact Mary learns when she learns the proposition expressed by "I am having an experience that feels like *this*" (where "this" names the reddish quale she has on first seeing an apple) might be the same fact as the fact she knows by knowing the proposition expressed by (say) "I am in brain state B." Now this seems highly implausible; it surely *seems* like these are just *different* facts, so that Van Gulick's suggestion has little intuitive force. But of course, he could reply that it might be that they *only* seem that way, just as it might seem, misleadingly, that the facts about water's freezing point and H2O's freezing point are different facts; and that in any case, mere appeals to the way things seem are hardly strong enough to save the knowledge argument as a *refutation* of physicalism.

The vague appeal to the way things seem is not all that's left to proponents of the knowledge argument, though. For we can appeal, as in the zombie argument, to the conceivability, and thus the *logical possibility*, of facts about experience and facts about brains states coming apart, which *shows* that the facts at issue *are not* one and the same fact. There is one important disadvantage of such an appeal, though: it appears to make the knowledge argument dependent on the zombie argument in such a way that there seems little point in bothering with the former, and weakens the hopes for a cumulative case against physicalism.

There is, however, something further to be said in favor of rejecting Van Gulick's reply, or at least keeping the burden of proof on him rather than Jackson. Think again of the mathematical example considered above: most people, including most philosophers, would no doubt take it to be wildly implausible to suppose that the proposition that 5 + 7 = 12 really is the same proposition as the proposition that 38 is the square root of 1,444, much less that the fact that 5 + 7 = 12 really is *one and the same fact* as the fact that 38 is the square root of 1,444. It is this sort of example which leads philosophers of language to suppose that taking propositions to be functions from possible worlds to truth values is simply an inadequate way to individuate them. These propositions just seem obviously to be *different* propositions, and the facts they are about seem just obviously to be *different* facts. And it is precisely our sense that this is so that leads us to adopt a more fine-grained scheme of individuating propositions in the first place. We don't suppose that this is necessary *merely* in order to take account of differences in concepts, but *also* because the propositions of which the concepts are constituents seem to be about *different facts*. But the suggestion that the facts that Mary learns on leaving the room are just the same facts as those she knew before seems just as intuitively implausible as the suggestion that these mathematical facts are the same. And if such an implausibility is, in the one case, *itself* precisely what gives rise to a more fine-grained account of mathematical propositions, so that it would be absurd to suppose that one could *defend* the claim that these are the same mathematical facts by *appealing* to a fine-

grained account, then it seems equally absurd and implausible to suppose that one could save physicalism from the knowledge argument by a parallel appeal to a fine-grained scheme of individuating propositions. In other words, it is, in part, precisely *because* we take physical facts and facts about qualia to be different sorts of fact that we find a fine-grained scheme of individuating propositions about them to be plausible in the first place. So it won't do to appeal to such a scheme in order to defend the claim that *aren't* different.

I conclude that this last sort of reply to the knowledge argument has no more force than the other two. That argument stands as compelling testimony to the reality of the qualia problem, and remains as formidable a challenge as ever to any physicalist attempt to solve it.

(d) Desperate measures: Levine's metaphysical necessity, Searle's biological naturalism, McGinn's new mysterianism

The second of what I have been presenting as the two most important attempts to show that physicalism cannot solve the qualia problem, namely the zombie argument, is in my view even more formidable than the knowledge argument. We have, in fact, already examined the main objection that has been made against it, which suggests a rejection of Hume's principle, and we've found it wanting. There are some other replies (or implied replies), however, which try more or less to solve the qualia problem within physicalistic boundaries; and they are eccentric enough that they reveal to just what lengths physicalism might have to go in order to avoid refutation.

As we saw earlier, Joseph Levine has suggested that even if there is an *explanatory* gap between physical facts and facts about qualia, this might not mean that there is a *metaphysical* gap between them. Now we've already seen why his approach is unconvincing if intended to show that the conceivability of a zombie world does not entail the logical possibility of such a world. But Levine also appears to hold that even if zombies *are* logically possible, physicalism could still be true; for perhaps there is a gap between what is *logically* possible and what is *metaphysically* possible. Or, to put the same suggestion another way, even if it isn't *logically* necessary that, given that all the physical facts obtain, the facts about qualia will also obtain, perhaps this is nevertheless *metaphysically* necessary. Maybe there is a kind of necessity (different from both physical necessity and logical necessity) which determines that even some worlds which are logically possible are nevertheless metaphysically impossible, and perhaps the zombie world is one such world.[4]

The rather glaring deficiency with this view is (as Chalmers has pointed out, 1996, p. 137) that it appears entirely *ad hoc*: there just seems to be no reason at all to accept the claim that what is logically possible might not be metaphysically possible, except that, if true, it would save physicalism.[5] And this is hardly an adequate reason

---

[4] Among positions which tend in the direction of physicalism, Levine's is the one which seems most clearly inclined toward this view, given his allowance of at least an explanatory gap between physical facts and facts about qualia; though as Chalmers notes, "few have explicitly taken this position in print" (1996, p. 371, n. 6). In any case, Levine is among those whom Chalmers cites as advocating the metaphysical necessity view in personal communication with him (Ibid.).

[5] Kripke's notion of *a posteriori* necessity is not germane here, because the view in question is intended to stand *independently* of Kripkean considerations, as a kind of "last ditch" defense of physicalism against the zombie argument; for as indicated in chapter 1, note 6, Kripke's work

to accept it, certainly not if one wants to appeal to it for the purposes of *defending* physicalism against objections.

Of course, the physicalist might respond by saying that without such a supposition, we would be left with a radical discontinuity in the natural order: everything else in the world seems to fit the physicalist model quite nicely, so how can it be that there's a *single* exception, conscious experience, which does not? How can it be true that human beings and other animals, all of which are ultimately as much the products of physical processes as everything else in the physical world, and for which have already have well-developed physicalistic theories of their other features, have a single feature which forever escapes physicalistic explanation? It is hard not to feel some sympathy with this reply, given the success the materialist or physicalist approach to the world has had in explaining so much of it. But it is nevertheless flawed on three counts. First, there is, at the end of the day, no *guarantee* that there couldn't be some single feature which escapes the physicalist story. If the evidence for this conclusion is overwhelming, it is idle stubbornly to close one's mind to the possibility and insist on any *ad hoc* device that might explain the evidence away. Instead, one might simply have to get to work and find out where exactly physicalism,

---

doesn't, despite appearances, give any comfort to the physicalist. To quote Chalmers (1996, p. 134): "[N]othing about Kripke's *a posteriori* necessity renders any logically possible worlds impossible. It simply tells us that some of them are misdescribed... One might have thought it possible *a priori* that water is XYZ, rather than H2O. In conceiving this, one imagines something like a world in which XYZ is the liquid found in oceans and lakes. However, Kripke's analysis shows us that due to the way the actual world turns out, we are misdescribing this world as one in which XYZ is water... Strictly speaking, it is a world in which XYZ is watery stuff. These considerations cannot show the impossibility of this apparently possible world; they simply show us the correct way to describe it."

otherwise so compelling, has gone wrong. Secondly, any assumption of merely a *single* apparent exception to the physicalist rule is an exaggeration at best: not only are there such other mental phenomena as intentionality which arguably defy materialistic explanation (though this is less evident than in the case of qualia, and any problem here might ultimately be derivative from the qualia problem itself), but there is also the vast ocean of mathematical truth which has since the dawn of philosophy been a thorn in the side of naturalism in all its forms. (I leave out here a reference to further problems that might be posed by arguments from philosophical theology, if only because, rightly or wrongly, there is far less consensus that there is any genuine subject matter here that is left out by the physicalist's ontology.) Finally, there is in fact a very good reason to *expect* that the facts about conscious experience should be, or appear to be, facts of a completely different sort from physical facts: namely that it is only *through* the former facts that we know any facts about the physical world at all, so that the former sort *should* seem to be different in kind from, and inexplicable in the same way as, the latter. What I mean is vague as I've just stated it, but it will become clear by the time we come to consider (in chapters 4-7) the two approaches to the qualia problem which really do get to the nub of the matter, two approaches which recognize that it's the fact that we know the world only *through* qualia which makes qualia a problem in the first place.

John Searle (1984, 1992, 1997), like Levine, is a philosopher inclined toward naturalism who is nevertheless sympathetic with the objections to physicalism we've

been looking at. But unlike Levine, he accepts not only an explanatory gap between physical facts and qualia, but a metaphysical gap as well: qualia are, in his view, irreducibly subjective, while physical facts are objective. Nevertheless, though he also, unlike Levine, rejects the materialist and physicalist labels, he denies being a dualist of any sort. He opts instead for a middle-ground position which he calls "biological naturalism" (1992, p. 1). But what middle ground could there be between these views?

Searle's answer is that an alternative to some variety of either materialism or dualism (or, for that matter, other traditional views like idealism) only seems impossible because of our tendency to use a philosophical vocabulary which biases us in the direction of assuming that these are the only possible alternatives. Terms like "mind" and "matter" are, he claims, typically *defined* in such a way that that assumption is only natural: "Thus we are supposed to believe that if something is mental, it cannot be physical; that if it is a matter of spirit, it cannot be a matter of matter; if it is immaterial, it cannot be material" (1992, p. 14). "But," he continues, "these views seem to me obviously false, given everything we know about neurobiology." Our starting point, Searle insists, should be the facts about consciousness we know from everyday experience and science: We know that consciousness is a subjective phenomenon and that physical systems like the brain are objective; but we also know that consciousness is a high-level feature of the brain, one that is caused by lower level features in the same way water's high-level feature of

liquidity is caused by lower-level features like the state of its molecules. And therefore we should conclude that consciousness is just a further *physical* feature of the brain which happens to be subjective, just as liquidity is a physical feature of water. These facts show that there is no special problem of seeing how consciousness can be a part of the physical world. Again, it is only the traditional terminology that makes us think otherwise:

> When I say that consciousness is a higher-level physical feature of the brain, the temptation is to hear that as meaning physical-as-opposed-to-mental, as meaning that consciousness should be described *only* in objective behavioral or neurophysiological terms. But what I really mean is consciousness *qua* consciousness, *qua* mental, *qua* subjective, *qua* qualitative, is *physical*, and physical *because* mental. All of which shows, I believe, the inadequacy of the traditional vocabulary. (1992, p. 15)

The solution of the long-standing mind-body problem is thus rather simple, on Searle's view. "Here it is: Mental phenomena are caused by neurophysiological processes in the brain and are themselves features of the brain" (1992, p. 1).[6]

Sympathetic as I am to Searle's suspicion that it is a mistake to think the traditional views in their various forms to be the only possible alternatives, I think his suggested solution to the qualia problem is no solution at all, but an attempt merely to define the problem away. Even if we grant that consciousness is in his sense

[6] Or, as he even more pithily summed it up in a talk I once heard him give: "Brains cause minds."

"physical," this doesn't solve anything given that it is also, as he acknowledges, irreducibly subjective, while the rest of the physical world is objective. For the relation between subjective and objective features is, as my statement of the qualia problem has made clear, *itself* what seems so difficult to account for. Searle insists that the objective properties of the brain cause the subjective features of consciousness, and that we know this from everyday experience. But this has never been primarily what is at issue; the problem is in explaining *how* this can be so, given that, as the knowledge and zombie arguments (which Searle accepts) show, there is no entailment from objective facts to subjective facts. And the emptiness of Searle's "solution" is only more painfully evident when he admits that "Biological naturalism raises a thousand questions of its own... [such as] *how exactly* do the elements of the neuroanatomy – neurons, synapses, synaptic clefts, receptors, mitochondria, glial cells, transmitter fluids, etc. – produce mental phenomena?" (1992, p. 1, emphasis mine). Elsewhere he cites as the most important problem facing the biological sciences the question: "How exactly do neurobiological processes in the brain cause consciousness?" and acknowledges that "we have only the foggiest idea of how it all works" (1997, pp. 3-4); in particular, "we don't have anything like a clear idea of how brain processes, which are publicly observable, objective phenomena, could cause anything as peculiar as inner, qualitative states of awareness or sentience, states which are in some sense 'private' to the possessor of the state" (1997, p. 8). Quite. But then, since explaining all *this* is precisely what is generally meant by the qualia problem or the hard problem

of consciousness, it is hard to see exactly what problem Searle thinks he has solved, or why he should treat solving the problem of *how* neural processes cause mental phenomena as if it were just a (very difficult, by his own admission) mop-up job that remains after the main problem has been solved. Given Searle's insistence on the equal irreducibility of objective and subjective features, I think it clear that, terminological fanfare notwithstanding, his view is really just another version of property dualism, as some of that view's representatives have themselves concluded (e.g. Nagel, 1995, p. 96; Chalmers, 1996, p. 370, note 2).[7] And as we'll see shortly, that view has serious problems of its own.

One final solution to the qualia problem must first be considered, however, though "solution" is perhaps not the best way to describe it. Colin McGinn, like Searle, rejects both dualism and all extant versions of materialism or physicalism; but like Levine, and unlike Searle, he thinks there is an unbridgeable explanatory gap between the physical and the mental. He thinks that consciousness is entirely the product of physical processes in the brain, and indeed, that there is a fully materialistic

[7] Searle gets at part of what, in my view, the actual solution to the qualia problem must involve when he suggests that an asymmetry exists between our knowledge of consciousness and our knowledge of everything else. He writes: "If consciousness is the rock-bottom epistemic basis for getting at reality, we cannot get at the reality of consciousness that way. (Alternative formulation: We cannot get at the reality of consciousness in the way that, using consciousness, we can get at the reality of other phenomena.)" (1992, pp. 96-97) and "There is, in short, no way for us to picture subjectivity as part of our world view because, so to speak, the subjectivity in question is the picturing" (p. 98). This is a bit vague, but it hints at something important I will try to make clearer later. In any case, Searle does not develop these suggestions in the sort of direction I think they inevitably lead; and one reason why he doesn't may be because of his rejection (as evidenced by his 1983, p. 58) of the indirect realist or representative theory of perception, which I think is implied by the facts about our access to the physical world to which he alludes, and which, as we'll see, holds the key to solving the qualia problem.

explanation of how this can be so; but he also thinks (unlike, perhaps, Levine, who appeals ultimately to the *ad hoc* semi-explanation of metaphysical necessity) that *we can never discover what this explanation is*. It is very likely, in McGinn's view (1991), that this explanation is cognitively closed to us, that we are so built that we simply lack, and forever will lack, the conceptual resources required to understand it, just as an armadillo must, given its cognitive limitations, forever be incapable of understanding arithmetic. Consciousness, even if a purely natural phenomenon, must forever remain a mystery to us, according to McGinn; for which reason Flanagan labels McGinn's position "the new mysterianism" (1992, p. 9) (the *old* mysterianism being dualism, which, unlike McGinn's view, takes consciousness to be inexplicable in scientific terms because it is a *non-natural* phenomenon).

Part of the evidence McGinn adduces in support of his position is, as might be expected, historical: the inability of philosophers to come to a consensus on a solution to the mind-body problem, despite centuries of effort, is just what we should expect if McGinn's thesis is true. But as McGinn acknowledges, this is hardly conclusive; maybe we just haven't found the solution *yet*. (Moreover, such an argument, if it supported McGinn's thesis, would support a host of other parallel pessimistic theses about other philosophical problems too; though this wouldn't necessarily deter McGinn, who, in later work (1993), has gone on to argue that many *other* traditional philosophical problems *also* have naturalistic solutions to which our minds are cognitively closed.) McGinn's main argument appeals rather to the very natures of the

two sides of the mind-body equation. What we'd need in order to be able to understand how material processes are responsible for producing consciousness is a grasp of some property, P, in virtue of which this occurs, in virtue of which there is a link between the brain and qualia. But there seems to be no way in principle of arriving at a concept of P, given that P would have to be a property which somehow *bridges* the divide between consciousness and the brain: introspection of consciousness itself will not yield the concept of P, for introspection can only ever reveal the one side by itself, and thus yield at most further concepts pertaining only to that side; nor could the concept of P be arrived at from perception of the brain, or reasoning on the basis of such perception, for the analogous reason that this latter process will only ever yield further concepts of the purely *physical* sort. In particular, McGinn says, it will always yield *spatial* concepts which by their very nature cannot apply to consciousness: it is inconceivable that we shall ever arrive at a concept, derived from perception of the brain, which will allow us to locate spatially another person's conscious experiences, and observe them as they're being produced by his brain. So P, even if it exists, is forever inaccessible to us, as is, consequently, any solution to the qualia problem.[8]

---

[8] One might raise the quibble that it *could* turn out that some third conceptual resource, apart from introspection or perception, might yield a grasp of P; perhaps the concept of P is acquirable *a priori* (and merely as yet unacquired) in the way Kant argued the concept of causation is. But of course, putting Humean scruples to one side, it can hardly be claimed that the link between qualia and the brain is as transparent as that between external causes and effects; nor does it seem likely at this juncture that we ever will arrive through mere conceptual investigation alone at a way of making it more transparent. (Nor, to bring Humean scruples back in, does the relationship between qualia and the brain seem as comprehensible as that between external causes and effects even *if* there is nothing

The relevance to materialism or physicalism, or at least to the naturalistic aspirations they are the best-known representatives of, is this: even if P is some purely natural property, that is, one that doesn't involve anything immaterial or supernatural like a Cartesian mental substance or an appeal to divine intervention, it will *still* be cognitively closed to us. So the fact that we cannot solve the mind-body problem does not entail that naturalism is false, or that we are forced to accept some form of dualism.

Now I think it must be acknowledged that, prima facie, what McGinn is suggesting *could* very well be true. There is simply no reason to suppose that everything that exists, even if entirely physical in its nature, *must* be comprehensible by us. But *why* should we suppose that we *are* in fact the situation he describes? That is, why suppose that consciousness *is* an entirely natural phenomenon, only we're unable to understand how? After all, the more natural interpretation of the situation we're left in by the knowledge and zombie arguments is that naturalism, as usually interpreted, should be rejected as *false*, not merely not fully comprehensible. The upshot of those arguments seems to be, not that qualia are physical, material, or natural properties *of an unusual sort*, but that they are *not properties of this sort at all*. They require, not that we give up on trying to *understand* naturalism, but that we give up *naturalism itself*, and adopt some form of dualism, most plausibly property dualism – this is certainly the lesson drawn by the best-known proponents of those

---

more to causation than constant conjunction. This is something to which we will return in the next

arguments, e.g. Jackson, Nagel, and Chalmers. McGinn's suggestion sounds, from the point of view of the critic of physicalism at least, entirely arbitrary, a case of special pleading, and a rather desperate last-ditch attempt to stave off the inevitable. It is as if a theist were to accept the claims that there are no good arguments for God's existence and that the problem of evil has no solution, but also insist that nevertheless, these facts constitute no good reason to adopt atheism, since it *could* be that theism is true and our minds are just cognitively closed to the theory that explains *how* it can be true. No naturalist would accept *that* as a plausible defense of theism; so why accept McGinn's position as a plausible defense of naturalism?

McGinn would no doubt defend his position as reasonable despite this objection on grounds similar to those on which, I suggested earlier, Levine might defend an appeal to the notion of a unique kind of metaphysical necessity, namely that it just seems implausible to suppose that there is a single phenomenon, qualia, which escapes the sort of physicalist account that prevails everywhere else. But what I said about such a defense in Levine's case applies no less to McGinn.

As we will see later on, there is in fact *some* truth to what McGinn says: there is indeed reason to suppose that the mind can never *fully* understand itself, at least not in respect to all of its features. But qualia are not among those inscrutable features. McGinn's position as a whole could recommend itself to us only if (a) there is no alternative way of dealing with qualia which can be considered broadly naturalistic,

---

section.)

and (b) there are decisive grounds for rejecting the property dualism that the qualia problem seems inexorably to push us towards. As we'll see in chapters 4-6, McGinn has failed to consider all the alternatives, and (a) is in fact false; even though, as we'll see in the next section, (b) is true.[9]

(e) Property dualism

Classical dualism, also referred to as *substance dualism* or *Cartesian dualism* (after Descartes, its best-known proponent) has it that minds are fundamentally different sorts of entities from material objects, that the mind is an immaterial substance which exists wholly independently of the brain. This view has been extremely unpopular among philosophers in the twentieth century, being seen as utterly incompatible with what modern science, especially evolutionary theory and neuroscience, have revealed about human nature (though even in recent years, it has not lacked able defenders, e.g. Foster, 1991, Hart, 1988, Popper and Eccles, 1977, Swinburne, 1986). Much more influential has been *property dualism*, which allows that the mind is identical with the brain, and that there are no immaterial *substances*, but insists that some mental *properties*, at least qualia, are themselves not *physical* properties of the brain. Property dualists acknowledge that the facts that mental

[9] Before leaving the discussion of physicalistic or at least broadly naturalistic attempts to solve the qualia problem, I should at least mention the recently popular strategy of trying to reduce qualia to intentional or representational properties, which are in turn reduced to physical properties of the brain, in more or less functionalist terms (Dretske, 1995, Tye, 1995). As Chalmers points out (1996, p. 377, note 38), these accounts appear to be subject to versions of the same sorts of objections made to standard functionalist accounts (e.g. the Chinese nation and inverted spectrum objections), objections I will discuss further in chapter 6, in the context of defending my own favored brand of functionalism, which, as will be seen, is of a *non-physicalist* variety.

functions of all sorts are so dependent upon neural processes, and to an increasing extent successfully explicable in neuroscientific terms, show that it is implausible to suppose the mind to be wholly distinct from the brain.[10] But they are also persuaded by such arguments as the knowledge and zombie arguments that qualia are peculiar features which in principle *cannot* be explained in neuroscientific terms or any other materialistic terms, so that they must be accepted to be *non-physical* features of reality. Some property dualists, like Jackson (1982) and Nagel (1974), more or less leave it at that, and hope that some future scientific and philosophical advances might make clearer exactly how these non-physical properties relate to the rest of the world, and why they appear in certain complex physical systems, i.e. brains. Their positions thus hardly count as true *solutions* to the qualia problem, but as, at best, clarifications of it, in that, if they are right, philosophers ought to stop approaching the problem by way of trying to characterize the facts about qualia in physicalistic terms. But other property dualists, like Chalmers (1996), try to develop more detailed constructive accounts, on which qualia are taken to be fundamental features of the universe (like space-time or electromagnetism) which are correlated with certain complex physical properties in a law-like way, where the laws governing them are in principle discoverable, so that a fleshed out picture of exactly how qualia relate with the rest of the world is possible. Indeed, Chalmers characterizes his approach as a kind of

[10] I lack the space to deal with substance dualism, which, though even less popular than property dualism, is worthy of an extended treatment. See Churchland, 1988, pp. 7-21, Flew, 1984, chapters 8 and 10, and Ryle, 1949, chapter 1, for some of the classic objections. Suffice it to say here that the

naturalism, *naturalistic dualism* in fact, since he takes his project to be that of merely *adding* to our conception of the basic features of the natural world and the fundamental laws governing them, rather than positing the existence of some vague supernatural realm which is unknowable, forever inaccessible to the normal methods of scientific inquiry. (He is, I should stress, thus quite self-consciously rejecting the standard construal of "naturalism" and its associations with physicalism or materialism. Nor does he have any illusions about solving the problem by redefining "physical" so as to include qualia (as Searle can be accused of having); he might better be thought of as suggesting a reconceptualization of naturalism so that both physical *and non-physical* features can be considered natural.) Chalmers-type property dualism thus counts as a full-fledged attempt to solve the qualia problem, one that, for obvious reasons, has none of the difficulties physicalist solutions do.

Property dualism in all its forms does have a very serious problem of its own, however, one as serious, in my view, as the problems facing physicalism. It is often accused of leading to *epiphenomenalism*, the view that mental states, or at least qualia, though perhaps caused by physical processes, have no causal effects in turn on the physical world. For if a person's neural processes, behavior, etc. (and even beliefs, desires, and the like, if a functionalist view of propositional attitudes, though not of qualia, is accepted), indeed, *all* the physical facts, would be exactly the same in a zombie world (where there are no qualia) as they are in the actual world, so that facts

---

difficulties for a property dualist solution to the qualia problem I will discuss would apply also to

about qualia are facts over and above the physical facts, then given that, as physics implies, the physical world is causally closed, it follows that qualia can play no role in bringing about physical effects in the actual world. Or in other words: if qualia did have physical effects in the actual world, then the physical facts would be different in the zombie world, since the qualia would not then be around to produce their effects; but the physical facts aren't different in the zombie world, so qualia can have no physical effects.[11] Now this would be a counterintuitive result, to be sure: it certainly *seems* as if qualia have causal effects on the physical world, e.g. it seems as if my sensation of pain causes such physical events as my pulling my hand away from the fire. But the clash with common sense is actually the least of the problems resulting from epiphenomenalism. The main difficulty is that if epiphenomenalism were true, then it seems that *we could have no knowledge of the existence of qualia.* This result is also contrary to common sense, but, more radically, it implies that property dualism is self-undermining, since if property dualism implies epiphenomenalism, and epiphenomenalism implies that no knowledge of qualia is possible, then property dualism (which insists on the existence of qualia as non-physical features of reality) implies about itself that if it were true, no one could possibly be justified in believing it!

The problem is that for us to know about qualia, it seems clear that they would have to have some sort of *effects* on us, just as we can know about tables, chairs,

---

substance dualism.

[11] Indeed, Jackson (1982) accepts this as a consequence of his position, and Chalmers, though not convinced it follows from property dualism, at least allows that it might, and that it would be a result he could live with (1996, p. 160).

rocks, trees, and the like only because *they* have effects on us. And if qualia are epiphenomenal, they *can't* have any effects on us. As Dennett says:

> Suppose, for instance, that Otto insists that he (for one) has epiphenomenal qualia. Why does he say this? Not because they have some effect on him, somehow guiding him or alerting him as he makes his avowals. By the very definition of epiphenomena..., Otto's heartfelt avowals that he has epiphenomena *could not* be evidence for himself or anyone else that he does have them, since he would be saying exactly the same thing even if he didn't have them. (1991, pp. 402-403)

Nor, as Dennett points out, would it help to suggest that Otto's evidence derives from introspection of, say, his beliefs about qualia. For if the beliefs he introspects are understood in physicalist or (standard) functionalist terms, then the qualia he claims to know about couldn't have any effects on *them* any more than on his behavior; while if those beliefs are also regarded as non-physical, then they themselves would be as epiphenomenal as the qualia are, and we're back where we started (1991, p. 403).[12]

One way for a property dualist to respond to all this would be to appeal to a Humean theory of causation, on which A causes B just in case A and B are constantly conjoined in experience, so as to defend the idea that there really are causal relations

---

[12] In fact we're stuck with the old *interaction problem* that faces substance dualism, of which the objection to property dualism we're looking at might be seen as a version. The idea is that if the mind is something wholly non-physical, then it seems it can have no causal connection to the body. For, since bodily behavior, being physical, appears to have physical causes entirely sufficient to account for it, there seems to be no room for an immaterial substance to have any causal influence on

between qualia and the physical world after all: pain is constantly conjoined with pulling one's hand away from the fire, so it can justifiably be said to cause the pulling away.

There are a number of problems with this approach, though, even apart from the controversy surrounding the adequacy of Hume's account of causation. For one thing, the reply seems question-begging: the property dualist would have first to know about the existence of qualia in order to establish that they are constantly conjoined with certain physical events, and whether he can know this is precisely what's at issue. But perhaps the property dualist can avoid this problem by arguing as follows: It certainly *seems* that there are qualia which cause behavior; that, at least, is uncontroversial. And the assumption that they really do cause qualia, in a Humean sense of "cause," allows us to explain why it seems that this is so – namely, because it really is so – without having to resort to either a physicalistic reductionism or eliminativism about qualia on the one hand, or an epiphenomenalist version of property dualism on the other, which all have problems of their own anyway.

But more serious problems loom. First of all, it isn't clear that epiphenomenalism can be avoided even on a Humean account of causation. Even if all physical causation is reinterpreted in Humean terms, it is still clear, given the zombie hypothesis, that the physical facts would be the same whether or not there are qualia, so that qualia can have no causal efficacy in the physical world. At any rate, in the

---

it; and furthermore, there seem to be conceptual problems with the notion of something wholly non-

face of the zombie hypothesis, a mere correlation between qualia and behavior couldn't establish otherwise: no plausible Humean account is going to allow just *any* constant conjunction to count as a causal relation, on pain of having to assert a causal relation between e.g. my arriving at 10:00 every Monday to teach the Phil 1 class and Jones's arriving at 10:00 every Monday to teach the badminton class.

Furthermore, there is another reason why mere constant conjunction between qualia and physical processes would fail to count even as a Humean causal connection, or to serve the explanatory purposes required by an interesting non-epiphenomenalist property dualism. For such a purported causal connection would not have the *fine-grained* character other causal connections do, even on a Humean interpretation. To take a favorite example of those who stress the inability of physicalism to account for qualia, the causal connection between the properties of H2O molecules and the liquidity of water is not merely a matter of a brute correlation between them: it's not that the state of the molecules causes the liquidity *in that, and only in that* such-and-such a state is always conjoined with liquidity. Rather, it's also that the fine-grained details of the molecules' state, their interactions with one another at room temperature, say, make it intelligible, indeed, even necessary, that the water they make up should be in a liquid state. Nor is it just examples of this sort of "bottom-up micro-macro" causation (as Searle calls it, 1992, p. 87) that exhibit such a fine-grained character. Even the stock one-billiard-ball-striking-another sort of case can be

---

physical (and thus non-spatial, lacking in mass, etc.) interacting with a physical system.

analyzed further, the precise relationship between the cause and effect fleshed out by reference to trajectories, environmental circumstances, the weight, precise shape, and even molecular structures of the balls, and so forth. Even if we ultimately work down to a level at which further analysis seems impossible, so that we're left, at that level, with brute Humean constant conjunction, we can still go very far beyond *merely* noting constant conjunction at the level of surface features. But though this seems to be true in every case with which the sciences deal, nothing close to it appears to be true or even conceivable in the case of the alleged causal relationship between qualia and the physical world. There we *do* seem to be unable to go beyond a brute correlation; and since this is untrue of cases we typically regard as genuinely causal, but generally *true* of cases we do *not* regard as genuinely causal (such as the correlation between my showing up at 10:00 and Jones's showing at 10:00), there is little reason to assume a Humean causal connection here rather than an accidental correlation – or at most an epiphenomenal relation.

Chalmers's preferred way of responding to the sort of problem raised by Dennett is not to insist on a genuine causal relationship of any sort between qualia and our beliefs about them, but rather to reject the assumption that knowledge of qualia must involve such a causal connection (1996, pp. 196-200). A causal theory of knowledge is "inappropriate" for understanding our knowledge of our own conscious states, he says, because on any causal theory, there is a gap between the knower and what is known which allows for skeptical hypotheses, cases where all our evidence for

something's existence is as good as possible, and yet that something still fails to exist. So, for instance, because our knowledge of physical objects is mediated by causal chains from the objects to our sense organs, and the usual causal connections can fail, so that what seems like good evidence for a belief in a physical object is really misleading (e.g. it might be hallucinatory), it follows that we can never be *certain* that any physical objects exist. But we *can* be certain that we are conscious. So our knowledge of our conscious states cannot be mediated by any sort of causal connection.

The first problem with this argument is that it shows at most only that causal connections are not *sufficient* for knowledge of conscious experiences, certain or otherwise; it doesn't show that they aren't *necessary*. Still, Chalmers might respond that the point is that *any* knowledge involving a causal connection *must* involve the possibility of error, so that if our knowledge of conscious experiences is certain, such connections simply can't figure into our knowledge of them at all.

But this brings us to the main problem, which is that there seems no non-question-begging reason to accept Chalmers's assumption that our knowledge of our conscious experiences is certain – after all, Dennett's point was precisely that property dualism itself seems to imply otherwise.[13] Now Chalmers claims that in fact only someone who already accepts a physicalistic reductionist or eliminativist view of qualia

[13] Note that taking our knowledge of qualia to be less than certain does *not* put us in the same position epiphenomenalism seems to put the property dualist in: even if we can go wrong with respect to our knowledge of our qualia, a causal link between qualia and our beliefs about them would still allow for

in the first place could take seriously the idea that we might be wrong about qualia; for if we might, for all we know, not have (irreducible) qualia even though it seems obvious that we do, we might as well just assume they don't exist, since this would make for a simpler worldview (1996, pp. 195-196). The obvious reply to this is that no one would take our lack of certainty about physical objects to imply that we might as well opt for an eliminativism or (phenomenalistic) reductionism about *them*; solipsism, however simple, is hardly to be recommended on that score. So why should the case be any different with consciousness? But Chalmers claims that there is a disanalogy here: In the case of physical objects, there is a gap between appearance and reality, in that it might *seem* that physical objects exist even if they do not; but in the case of consciousness, there is no such gap, for the "seeming" itself just is a conscious experience (1996, p. 195). And perhaps the implicit reply to my objection here would be: We're certain we're conscious, since there's no gap between appearance and reality where it's concerned; and the zombie argument, etc. shows that property dualism is true. So there must be *some* error in Dennett's argument to the effect that property dualism and certainty about consciousness are incompatible. This reply fails, though, because Chalmers fails to take account of the fact that there is a non-experiential, *cognitive* sense of "seeming" as well, as in "At first it seemed to Smith that Chalmers's arguments were sound, but on further reflection he concluded that they were not." And with *that* sense in mind, we can see that there could indeed be a

---

(defeasible) epistemic access to them; whereas if there is, and can be, no causal connection *at all*, it is

gap between appearance and reality even in the case of consciousness, where it could seem (cognitively) that one was having conscious experiences even when one was not. (This is a result that Chalmers, of all property dualists, is going to have a difficult time avoiding, considering that he allows for a functionalist account of the propositional attitudes (though not, of course, of qualia) and accepts that qualia are explanatorily irrelevant to our judgements about our conscious experiences (1996, p. 192).)

Of course, Chalmers is right to want to *try* to resist the conclusion that our knowledge of conscious experiences isn't certain, counterintuitive as that conclusion is. But property dualism itself seems to invite this result more than its rivals do. After all, if qualia were just physical properties of the brain, the way *might* be opened to an account on which, in one's knowledge of his own qualia, the *state* of knowing and the *object* of knowledge were identical, so that there would be no metaphysical gap giving rise to an epistemological gap between appearance and reality. Such a suggestion would require development, of course; but the point is that making qualia out to be fundamentally the same sort of properties as beliefs, neural properties, and physical properties in general would at least allow for the possibility. But on a property dualist account – at least on Chalmers's version, where beliefs, desires, and the like are allowed to be functional properties of a physical system – qualia would be fundamentally different sorts of thing from the brain, and even from the beliefs one has

---

hard to see how there can be *any* justification for our beliefs about them, even of a fallible sort.

about them, so that a metaphysical gap, and thus, plausibly, an epistemological gap, seems inevitable.

Chalmers has one more arrow in his quiver, however, namely an alternative account of our knowledge of conscious experiences, one which does not appeal to any kind of causal connection. If this account succeeds, and, more to the point, shows how knowledge of qualia can be certain, the objections we've been developing will be moot. On his account, the mere *having* of an experience confers justification on the belief that one has it, without mediation of any sort of causal link (1996, p. 196). There is, he says, "something intrinsically epistemic about experience. To have an experience is automatically to stand in some sort of intimate epistemic relation to the experience – a relation that we might call 'acquaintance'" (1996, pp. 196-197). There is, that is to say, a *conceptual connection* between having an experience and being in this sort of epistemic relationship to it. Thus there is no need for any causal connection between qualia and our beliefs about them: we are in *direct* epistemic contact with them – and moreover, this direct contact provides us with *certainty* regarding their existence.

Now it might be thought that Chalmers is here appealing to the old idea of *incorrigibility*, on which our judgements about our experiences are infallible; but he explicitly rules out such an interpretation, and allows that there can be unjustified beliefs about experiences, e.g. when one is distracted (1996, p. 197). But then it's difficult to see exactly how simply having an experience could give one *certainty* about

its existence: if I can, after all, be wrong about my experiences in *some* cases, why not in all of them? Perhaps the claim is that we can go wrong about our experiences in a minor way, e.g. by failing to notice a red patch in a portion of the visual field (say, the corner of one's eye) and thus forming the belief that there are no red objects (and so no red qualia) before one; but not in a major way: this might be the force of Chalmers's claim that one can, if he seems to be having normal everyday experiences, *know* that he is not *really* having experiences of "a host of bright flashing yellow and green experiences with a deafening noise, say" (1996, p. 195). But given the possibility of the minor errors, it is hard to see how it can be denied that more radical errors are at least *possible*. Many traditional arguments for external world skepticism, after all, rest on premises about relatively minor or uncommon perceptual errors, such as illusions and hallucinations, the idea being that there is no non-arbitrary way of ruling out error in *all* cases if one acknowledges it in *some*. So why should things be any different in the case of knowledge of one's experiences? Indeed, there is at least *some* empirical evidence supporting the possibility of radical errors about one's experiences, such as the confused experiences involved in cases of synaesthesia, where patients report "seeing noises," "hearing colors," and the like.

Alternatively, Chalmers's position might be that one can go wrong with respect to the precise *character* of one's qualia, but not with respect to their *existence*, appealing to an apparently much clearer distinction than that between major and minor errors in judgements about experience. But then it's hard to see exactly what is being

ruled out, given that qualia *just are* the way things seem, subjectively, to be: if I can be mistaken about whether the qualia I'm having are red or green or blue – or *color* qualia at all, for that matter, as opposed to auditory, tactile, or some other kind of qualia – then it isn't clear *what* it is I can be certain exists. ("I don't know whether it is something bright red, or rather like a dull pain, or perhaps instead a pungent odor... but *it* is real all right!")

In any case, it isn't clear why, on Chalmers's account any more than on a causal account, we should rule out the possibility, already alluded to, that things might *cognitively* seem to be a way they are not in the case of beliefs about experience no less than in the case of beliefs about the external world. I might *believe* I'm having an experience even though I'm not, my reason being that it *seems* (cognitively) that I am; so perhaps I'm *always* so deluded. Maybe I really am a zombie, even though I cannot seriously believe that I am any more than I can bring myself to believe any other skeptical hypothesis.[14]

Chalmers says that this sort of objection falsely supposes that a situation in which my beliefs are as they are now, but they're all false, is a situation which is *evidentially* identical to my actual situation; whereas in fact, my evidence is "more

[14] Descartes' famous "evil demon" skeptical thought experiment adds some support to this possibility, since on some interpretations, the demon could be toying with my cognitive processes as well as my experiences; though in the zombie case, of course, it is supposed that the zombie's cognitive processes (understood, again, in purely physicalistic functionalist terms, and apart from the qualia associated with cognitive processes in our own case) are normal, so that any argument for the claim that zombies could go wrong about their experiences couldn't rest on the idea of malfunctioning cognitive processes. But of course, such an argument needn't rest on this anyway, but rather on the idea that

primitive" than my beliefs, so that a situation in which my beliefs are all the same might nevertheless be a situation in which my evidence is different (1996, p. 199). But this misses the point: of course my evidence for my beliefs about my experiences might be as I suppose it to be, but the point at issue is really how I can *know* whether or not it is. If Chalmers says "I may not know whether or not I'm hallucinating that chair, but at least I know I'm having the qualia associated with chairs," we can still ask "*How* do you know even *that*?" "Well, it just *seems* like I do, and it's seeming that way is all the justification I need!" But a zombie would believe the same thing! "But *I* have evidence the zombie doesn't have!" Chalmers would retort. But that's what the zombie believes too, because it also seems (cognitively) to *him* that he has such evidence. Any response Chalmers could give to such questions would just invite further questions about whether he *really* has the evidence he thinks he does, or only *seems* to. And ultimately, that puts him in the following dilemma: if he says that he "seems" to in the *cognitive* sense of "seems," then he's saying something even a zombie would believe; and if he says, even to himself, that he "seems" to in the phenomenal, qualia-involving sense of "seems," then he's begging the question.[15]

The upshot of this is that there seems to be no way a property dualist (and *maybe* anyone *else*, for that matter) can give a non-question-begging defense of the

---

it's logically possible that even perfectly functioning cognitive processes could exist in the absence of qualia.

[15] We might also note that Chalmers's argument here appears not to sit well with his argument for a nonreductive version of functionalism, which we'll be examining later, in which he acknowledges that it is *logically* possible, though not, he thinks, empirically possible, that one's judgements about his experiences could be systematically in error (1996, chapter 7).

claim that our knowledge of qualia is certain; and this means that there is no basis for rejecting the assumption, plausible in any case, that our knowledge of qualia, like most, if not all, of our other knowledge, must involve a causal connection.[16] But then, if property dualism entails epiphenomenalism, as it appears to, it also entails that we can have no knowledge of qualia; in which case, property dualism is false.

(f) A Kantian antinomy?

Some physicalists, like Dennett, take the difficulties with property dualism we've been looking at to be evidence that there must be a flaw in the arguments against physicalism, so that if property dualism is to be rejected, physicalism "wins" by default. But this is, in my view, too glib. There's certainly nothing in what was said in the last section that points to a *specific* error in either the knowledge argument or the zombie argument. The problems with property dualism seem to me to be no *more* serious than the problems with physicalism, and in fact the two views appear to be at a standoff. Indeed, the common sense presuppositions we bring to bear on these issues arguably make this more or less inevitable. Our everyday conception of the material world seems to leave no place in it for qualia, and consciousness, as we know it in introspection, seems too ethereal to have any sort of interaction with those bulky,

[16] It might be suggested that mathematical knowledge is a counterexample to this, since we arguably have knowledge of abstract mathematical objects even though it seems impossible that there could be a causal connection between us and them (though of course many philosophers, insisting on the need for a causal connection, would argue that this implies that we *don't* in fact have knowledge of abstract mathematical objects, and that mathematical knowledge shouldn't be construed in such Platonistic terms in the first place). But it is the *abstract* nature of mathematical objects which makes the notion of a causal interaction with them problematic, and qualia don't have this abstract character, but are

fleshy things we call our bodies; and there just seems to be no way to reconcile these intuitions, as *centuries* of inconclusive argument indicates.

What we're left with, I suggest, is something not unlike one of Kant's *antinomies*, where reason leads us, by arguments equally persuasive, into an apparent paradox. And like some of the Kantian antinomies, this one can only be resolved by rethinking some of the assumptions we take for granted in our most general thinking about the world. In particular, it must involve, first, a rejection of the largely unacknowledged assumption made by both physicalists and their opponents that we have a transparent grasp of the intrinsic nature of matter. As the Russellian position we shall examine in chapters 4 and 5 makes clear, seen in the light of the true character of our knowledge of the external, physical world, the qualia problem takes on a whole new complexion. But even Russellians make a parallel assumption we shall also find reason to reassess in chapter 6, namely that we at least have a transparent grasp of the nature of *mind*, and this assumption leaves even the Russellian position with deep inadequacies of its own. Solving the qualia problem, I will argue, will require no less than a radical reassessment of *both* of these assumptions, assumptions which go deep in our commonsense picture of the world.

(in this respect, anyway) more like the concrete things our knowledge of which does seem essentially to involve a causal connection.

4. Russell and the identity theory

(a) Our knowledge of the external world

I've said that the usual approaches to the qualia problem all share certain assumptions – assumptions which they inherit from common sense, but which are, nevertheless, false, and block the way to a genuine solution to the problem. One of those assumptions is that we have a transparent understanding of the nature of the material world – transparent enough, at any rate, to enable one to say (if one is a dualist) that mind and matter cannot possibly be of the same sort, or (if one is a materialist) that first-person, subjective qualities cannot possibly be properties of the physical world. This is an assumption rejected by an unusual and neglected approach to the mind-body problem associated with Bertrand Russell – an approach which, however, still counts as an (idiosyncratic) version of the mind-brain identity theory.

Russell is not the only philosopher to take the approach we're now going to be examining, nor even the first: Moritz Schlick (1985) was probably its earliest 20th century advocate, and there's reason to believe that it has antecedents in Kant, Schopenhauer, Wilhelm Wundt, and W.K. Clifford (see Lockwood, 1989, pp. 169-171); and other 20th century advocates and sympathizers include Herbert Feigl (1967), Grover Maxwell (1978), Michael Lockwood (1989), David Chalmers (1996), and Galen Strawson (forthcoming). But Russell has become the best-known advocate of this position (perhaps because he presented and argued for it in a number of works ranging over a thirty year period, from 1927's The Analysis of Matter (reprinted

1954), to Human Knowledge (1948), "Mind and Matter" (1956), and My Philosophical Development (1959) – and it even gets a mention in A History of Western Philosophy (1945, pp. 833-834)), and recent discussions of it almost always characterize it as the "Russellian" view. This is one reason why I will characterize the view under consideration as Russellian; another, more substantial reason, is that Russell was perhaps the first to develop it in the manner I find the clearest and most compelling, by proceeding from a defense of the *indirect realist* theory of perception, and I will follow his example in my exposition.[1]

Indirect realism is, roughly, the view that in perceptual experience, we are (as common sense supposes) indeed made aware of objects and events existing external to the mind (hence it is a form of realism), but (contrary to common sense) never *directly* so – our awareness of them is always mediated by our direct awareness of something internal to the mind, say sense-data or qualia (hence the modifier "indirect").[2] Our perceptual situation, on the indirect realist view, is thus analogous to the situation of someone watching a person being interviewed on a live television broadcast: the

[1] It should be noted that, as on many other issues, Russell's views on perception changed during the course of his long career. A commitment to indirect realism is, generally speaking, not to be found in works prior to the ones just mentioned. I'll say a little more about the evolution of Russell's views vis a vis the issues at hand later on.

[2] This view is also sometimes called "causal realism" (since it holds that the external objects which we are aware of indirectly are the causes of the entities we are directly aware of in perceiving them), and "representative realism" or the "representative theory" (since it is often said that the internal objects we are directly aware of "represent" the external objects we perceive indirectly). But I prefer the "indirect realist" label, since there are other views sometimes classified as causal theories of perception which nevertheless reject the notion that we are never directly aware of external objects, and since I think it is at best misleading and at worst false to speak of what we are directly aware of as "representing" external objects, since, for reasons we'll see, there are no grounds for thinking that external objects are in themselves the way they appear to us in perceptual experience.

viewer is aware of the person being interviewed, but only indirectly, via his direct awareness of the image on the television screen – though of course, on the indirect realist view, the perceiver in this case isn't directly aware even of the television image itself, but only of the qualia produced in his mind by whatever brain state is caused by the light from the image striking his retina, etc. (The *direct* realist view, often said to be the commonsense view, would then be the view that in perceptual experience we are directly aware of external objects, in a manner analogous to our apparently direct awareness of a person when he's right in front of us, rather than in a television studio miles away.)

The best-known argument for this position is the *argument from illusion* (perhaps most famously defended in this century in Ayer, 1940), which goes more or less as follows: Perceptual experiences we take to be veridical (that is, as presenting a reliable picture of the external world), such as seeing a bent stick, are intrinsically indistinguishable from those which are misleading, such as those involving the illusion of a straight stick's appearing to be bent when submerged in water – there is no way to tell, from the character of the experiences considered by themselves, whether or not they are trustworthy. And this supports the notion that whatever one is directly aware of in the one case must be something of the same sort as what one is directly aware of in the other, since otherwise, it seems plausible to suppose, there would be some difference in the intrinsic character of the experiences. But in a case of perceptual illusion, one cannot be directly aware of an external, physical object: in the stick-

submerged-in-water case, for instance, one cannot be directly aware of a bent stick, because there *is* no bent stick there. And so, neither can one be aware of external, physical objects in cases of veridical perception. But then what one *is* directly aware of must be something else – and the most plausible candidate for this something else is something mental, the *experience* itself, in particular the qualia which characterize the experience.

Despite having a long history in the philosophy of perception, this sort of argument has in recent decades come in for heavy attack, most famously in J.L. Austin's Sense and Sensibilia (1962). Austin argues that the typical examples used by proponents of the argument from illusion are simply misdescribed: for example, "it is simply not true to say... that seeing a stick refracted in water is exactly like seeing a bent stick" (1962, p. 49). It is instead like seeing a *straight* stick *submerged* in water – indeed, it *is* seeing a straight stick submerged in water! So it is false to say that veridical perceptions and illusions are indistinguishable. So there's no basis for the claim that in this sort of case one isn't directly aware of an external physical object; and thus the inference to the claim that we never directly see external objects is blocked.

While I'm not convinced that cases of illusion lend *no* support to indirect realism, I won't try to defend that sort of argument here. I want instead to focus on another argument for indirect realism, an argument which appeals to considerations of a sort Russell himself thought the most compelling. As Austin also points out, the

"argument from illusion" label is often applied to arguments which appeal, not to illusions of the sort just considered, but rather to *hallucinations*, even though these are, strictly speaking, very different from illusions. And though Austin no doubt wouldn't have concurred, arguments from hallucination are, in my view, much more formidable than arguments from "illusion," as that term is usually understood – especially when conjoined with considerations about *causation*, in particular the fact that there is a causal connection between a table, say, and one's "table-ish" experience of it, the existence of which implies a distinction between the two and makes plausible the suggestion that one is only aware of the former via awareness of the latter.

Now while various philosophers have put forward distinct arguments from hallucination and causation for indirect realism, Howard Robinson has suggested, rightly in my view, that the clearest and most effective case is made by a single argument which brings both considerations together. Robinson's own formulation of such an argument is as follows (1994, p. 151):

> (1) It is theoretically possible by activating some brain process which is involved in a particular type of perception to cause an hallucination which exactly resembles that perception in its subjective character.
>
> (2) It is necessary to give the same account of both hallucinating and perceptual experience when they have the same neural cause. Thus, it is not, for example, plausible to say that the hallucinatory experience involves a mental image or sense-datum, but that the perception does not, if the two have the same proximate – that is, neural – cause.
>
> (3) Therefore, perceptual processes in the brain produce some object of awareness which cannot be identified with any feature of the external world – that is, they produce a sense-datum.

Premise (1) appeals to the notion that perceptual experience is the end result of a chain of causal processes that come between the object perceived and the experience itself, the penultimate link in the chain being some brain process or other. In the ordinary case, the causal chain goes (to put things very crudely) something like this: light strikes a table, and is reflected off of it and strikes the retina, whereupon the optic nerve sends signals to more central areas of the brain, where various neural subsystems then process the signals further until forwarding them to the optical centers, neural processes within which finally bring about a "table-ish" visual experience. Given that there is some brain process which serves as the last causal link before the experience is produced, it seems possible at least in principle that that process could be brought about in a way that is out of the ordinary, through electrical stimulation of the brain, say, without there being a table, reflected light, etc., at all – in which case an experience would be produced which is in its first-person, qualitative character exactly like the ordinary experience, but which due to its deviant causal ancestry (and especially the absence of a table) would be a *hallucinatory* rather than a veridical experience.

Premise (2) rests on the principle "same proximate cause, same immediate effect" (Robinson, 1994, p. 154). If the same sort of neural process is the cause of both the veridical and hallucinatory experiences, then the effects must be of the same sort as well. Thus, if in the hallucinatory case that process causes an experience the direct object of awareness in which is a constellation of sense-data or qualia – as it

certainly appears to, since there's nothing else there to be aware of, directly or indirectly – then it must cause the same thing in the veridical case. And it follows, then, that even in veridical perception, one is only ever directly aware of qualia, and aware of external physical objects only indirectly, when they happen to be the distal causes of the qualia in question.

Now as Robinson notes, the claim made in premise (1) is plausible enough that it is rarely challenged by critics of this sort of argument. The most obvious way to try to circumvent it would be to argue for some kind of eliminativism about experience, or at least about what is said to be common to the veridical and hallucinatory experiences, namely the qualia involved; but as I've already tried to show, this sort of eliminativism is implausible (and certainly not *more* plausible than indirect realism). The real controversy is over premise (2).

One way to challenge (2) would be to challenge the principle that the same proximate cause must produce the same immediate effect, a principle which, though not a necessary truth, still has a considerable amount of empirical plausibility. Robinson rejects, rightly in my view, the suggestion that even if this principle applies in cases where both cause and effect are physical, there is no reason to suppose it applies to cases where the cause is physical and the effect mental. One defense of this suggestion would be to appeal to the fact that while in cases where both cause and effect are physical, we have ways of identifying the effect *qua* the kind of effect it is which are independent of the cause (so that we can in these cases establish the truth of

the principle at issue), we have no such independent identification of the effect where that effect is mental: our reasons for calling one experience veridical and another hallucinatory are precisely that their (distal at least, if not proximate) causes are different (so that there is no way to establish that the principle applies to the physical-to-mental case). So perhaps in the latter sorts of case, the effects can be different; in particular, maybe the effect of the sort of brain process in question is direct awareness of a constellation of qualia in the hallucinatory case, but direct awareness of a physical object in the veridical case. As Robinson notes (1994, pp. 157-158), the implausibility of such a disjunctive analysis lies in the fact that it makes a complete mystery of the phenomenon of hallucination: How exactly does the relevant brain process "know" when to produce direct awareness of a physical object and when to produce (hallucinatory) direct awareness of qualia?[3] *Why* exactly would it have the capacity of doing the latter at all? On the indirect realist view, by contrast, hallucination is intelligible: the relevant brain process does the *same* thing in *every* case, namely produces direct awareness of qualia, and a hallucination occurs when this is not associated with certain other factors which typically accompany it. I would add, moreover, that even the putting forward of this response implicitly concedes too much to the proponent of the disjunctive analysis, namely that there really is any special need for the indirect realist to justify applying the principle in question to the physical-to-

[3] More precisely: What feature of the brain process, considered by itself, could be responsible for causing it to produce direct awareness of a physical object under some circumstances, and direct awareness of qualia under other circumstances?

mental case. For, first, I should think that the burden of proof is on anyone who wants to *deny* that the principle "same proximate cause, same immediate effect" applies to this sort of case. Given that there are causal connections between brain states and mental states, as is almost universally acknowledged, why suppose that these are any less governed by this principle than any other sort of case? The uniformity of nature would seem to require assuming that they are so governed until we have reason to believe otherwise.[4] Second, as physicalists never tire of reminding us, and as the Russellian would, as we'll see, agree (though on a very different interpretation of the claim), there are grounds for holding that mental states are themselves (certain kinds of) physical states, so that if the principle in question is true of physical-to-physical cases in general, it follows that it is true of physical-to-mental cases as well. (Though I grant that anyone who follows the Russellian line must be careful here, lest this second point be presented in a question-begging way: the Russellian *defense*, against the standard objections, of an identity between mental and physical – though not the *arguments* for the identity themselves – appeals, in part, precisely to indirect realism.)

A more common way to try to undermine (2), and indirect realism itself, however, is to challenge its implicit commitment to an *act/object* analysis of perception. Indirect realism takes the direct object of the act of perception, whether

---

[4] Might such a reason be provided by the problems we saw the best-known attempts to solve the qualia problem have in spelling out precisely the relationships (including the causal relationships) existing between brain states and mental states? No. Justifying the principle "same proximate cause, same immediate effect" in any domain would seem to require attention only to one particular feature of causal relationships, namely the general correlation (what Hume would call the "constant

veridical or hallucinatory, always to be a quale or a constellation of qualia: But why suppose there *must* always be a perceived *object*, an *entity* of some sort, so that if, in hallucination, it is not a physical object, it must be some other kind of object? Perhaps the correct thing to say is, not: "It seems that I am directly perceiving a physical object, but I'm really only perceiving a quale," but rather just: "It seems that I am directly perceiving a physical object, but I am not," where the "seeming" is not itself thought to be any sort of object. And one way to cash this suggestion out would be to adopt what is known as an *adverbial* analysis, where the correct way to describe an hallucinatory experience of a red object is, not: "I am directly aware of a red quale," but rather something like: "I am being appeared to redly," where the (admittedly strange sounding) neologism "redly" allows us to describe the experience without committing ourselves to the existence of any sort of mental object.[5]

Now the strangeness of this way of talking has itself caused many philosophers to be suspicious of this strategy. And, more substantially, a common objection to an

---

conjunction") between a cause and its effect, and establishing such correlations is the least problematic part of accounting for the causal relationships between the physical and the mental.

[5] John McDowell (1986) suggests that cases like that in which someone says (out loud or to himself) "That man over there is Smith," where there is no man over there but only a shadow, say, involve "cognitive illusions" such that the person isn't *really* even thinking the thought that *that man is Smith*, even though it seems that he is. (McDowell's grounds for such a view involve a commitment to the existence of "Russellian propositions" – yet another interesting idea we owe to Bertrand Russell – which are propositions of which particular objects of thought, such as particular people, mountains, galaxies, etc. etc., are themselves constituents. If there *is* no particular man over there, then there simply can be no Russellian proposition about the man to serve as the content of the thought, and thus even though the person thought he was thinking such a thought, he couldn't have been.) McDowell is no friend of indirect realism, and this sort of idea parallels, and may suggest a way of giving content to, the idea that in *seeming* to aware of an object, one isn't *really* aware of anything, not even something mental; though McDowell isn't necessarily committed to the adverbial analysis, and Robinson (1994, p. 247) classifies him as a proponent of the disjunctive analysis considered earlier.

adverbial analysis is that it is hard to see how it might deal with even slightly more complex examples: "I am being appeared to squarely and redly" is ambiguous, since it isn't clear whether what is being described is an experience of a red square or an experience of redness and squareness (where the redness and squarenes aren't instantiated at the same place in one's visual field). But the question of the prospects for fleshing out an adverbial analysis is moot, in my view, for there is ample reason to prefer the act/object analysis to it in any case.

Consider the example of mirror images. An image in a mirror cannot, I think, be an objective object of perception. For consider any particular spot *s* on the surface of a mirror (marked perhaps with a small black dot), two observers, *A* and *B*, both of whom are looking at the mirror from different points in space, and two physical objects, *O1* and *O2*, positioned at different points in space between the observers on the one hand and the mirror on the other. As the reader can easily verify, it is possible for *A*, *B*, *O1*, and *O2* to be so situated that what *A* sees at point *s* is the image of *O1*, say, while what *B* sees at the very same point, at the very same time, is the image of *O2*. So the mirror images of *O1* and *O2* occupy the very same point in space at the same time. But of course, no two physical objects can possibly occupy the same point in space at the same time. So, these two images, insofar as they are objects at all, are not *physical* objects.

Of course, this is just the sort of example a proponent of the argument from illusion might appeal to, and I suspect that it would be far more useful to him than the

usual submerged stick sort of example is. But the use I want to make of it here is slightly different. Those who object to the act/object analysis deny that there is *any* object of experience in the hallucinatory case – but, especially given that their aim is to defend direct realism, they'd have to admit that there *are* such objects in the veridical case, namely at least physical objects. But then the mirror example shows, I think, that the objects of perception in the veridical case are not limited to physical objects. For my perception of a mirror image is a veridical perception all right – I'm not *hallucinating* the image, after all – but as we've seen, it isn't perception of a physical object (though of course, in perceiving it, I might also perceive, indirectly, the physical object of which it is an image). It won't do to say I only *seem* to see something, where the "seeming" isn't itself an object – for I *really do* see the image (and even if the image itself isn't in the relevant sense "out there," there are objective correlates which are – photons striking the mirror a certain way, etc. – and which can in principle be detected by third parties, unlike in the hallucination case). Nor does it seem plausible to suggest that an adverbial analysis might be applied here as in the hallucinatory case: for in the very same experience in which I see the image, I clearly really do perceive certain physical objects, i.e. the mirror itself, the wall next to it, etc., and there is a "smoothness" to the experience such that the "image-ish" part of it blends quite imperceptibly into the "mirror-ish" and "wall-ish" parts of it; so it seems highly arbitrary and implausible to suppose that the experience can be partitioned into portions which are susceptible of an adverbial analysis and/or which do not involve an

object of perception, and portions which are not. But what *am* I aware of in being aware of a mirror image? Well, the most obvious answer is: a constellation of *qualia* of the sort usually associated with perception of the physical object of which the image is an image. That is, I'm aware of something *mental*. Thus, we seem led ineluctably to the conclusion that the direct object of perception in at least some cases of veridical perception is a *mental* object – and this certainly takes the wind out of the sails of critics of the act/object analysis, who implicitly suppose a clear-cut distinction between cases where objects of perception must be admitted, but only physical objects, and cases where *no* object, physical or mental, need be admitted.[6]

Further support for the act/object analysis – or at least, more directly, for indirect realism over direct realism – is provided by a variation on some of the traditional thought experiments often used to characterize skepticism about the external world. Think in particular of the "brain in a vat" scenario. We are to imagine a brain kept alive, not in a human skull, but in a vat of nutrients, and receiving constant stimulation, via electrodes attached to it, from a supercomputer such that its experiences are exactly the ones a normal person might have. The brain might think, just as I do, that it is sitting before a computer and typing on a keyboard, but unlike me, it is in reality floating in the vat in a lab somewhere. Now the standard use to which this is all put is to set up the skeptical challenge: "How can you possibly

[6] A similar argument could be developed which made reference instead to rainbows (another favorite example of proponents of the argument from illusion) – perceptions of which are clearly veridical even though, given that whether one sees a rainbow at a particular spot in the sky (and thus whether it exists there at all) depends on one's point of view, they seem hardly to be physical objects.

rationally justify the belief that you aren't a brain in a vat?" But what I want to do here instead is to alter the thought experiment slightly so that the brain is not, as is typically imagined, *completely* deluded. Suppose that there are indeed, contrary to what the brain supposes, no keyboards, tables, chairs, rocks, trees, and the like for miles; but that there are at least some rather less interesting objects – bland middle-sized geometrical figures such as cubes, spheres, pyramids, and the like – scattered about the portion of the lab in which the brain sits in its vat. And suppose that there are some crude photoreceptor cells hooked up to the computer system feeding the brain its experiences which can register any changes in these objects and feed the information into the system. Then we can imagine that the computer takes account of this information in such a way that, even though the brain's experiences are just like mine (let us continue supposing), it is nevertheless not *completely* cut off from its environment because its experiences of changes in the position of (what it takes to be) the chair on the opposite side of the room, say, co-vary in a regular way with changes in a certain cube in the lab. The computer still "fills in" *most* of the details of the brain's perceptual experience, but at least some of those details are determined by the information fed to the system via the photoreceptor cells.

We can, I think, quite plausibly say that the brain has some – extremely limited – veridical perceptual experience of its environment. For example, we can suppose that the experiences it has with chairs, at least, are (partially) veridical. And the lesson I want to draw is this: If we say that in its veridical experiences the brain is directly

aware of objects – as even critics of the act/object analysis and of indirect realism in general would say of *our* everyday veridical experiences – they *can't* plausibly be said to be *physical* objects. For what the brain seems *directly* to be aware of is e.g. a "chair-ish" thing, while the physical object corresponding to the experience is a *cubical* thing. It clearly seems to be aware of the cube only *indirectly*, via its direct awareness of something mental, namely its "chair-ish" qualia. But *our* awareness of chairs, tables, etc., given that it is also mediated by a causal chain (even though one which – we assume! – doesn't include supercomputers and photoreceptors) differs from the brain's awareness of its external environment only in degree. (Moreover, since physics tells us that physical objects are composed of tasteless, odorless, colorless, and generally unobservable particles separated from one another to such an extent that any given object is mostly empty space, etc., we can hardly be said to be *that* much closer to the truth about them than the brain in the vat is! And as we shall see shortly, there is reason to think that what we do know about the physical world is, relatively speaking, very little indeed.) So there seems no reason to suppose that our perception of the external world is any less mediated by *direct* awareness of qualia than the brain's is.

This quite naturally brings us to the *motivation* so many philosophers have had for trying to avoid indirect realism, even though their *reasons* for rejecting it seem, under scrutiny, to be arbitrary and *ad hoc*. Indirect realism, it is widely thought, threatens us with skepticism about the external world: if all we are ever directly aware

of is our own qualia, or with, as it has traditionally been described, a "veil of perceptions," how can we ever be *justified* in believing, as we all do, that there is a real world of physical objects existing beyond that veil? If skepticism can't quite be *proved* to be false – as probably most philosophers today would concede – so that the alleged tendency of indirect realism to lead to skepticism cannot arm its critics with a *reductio ad absurdum* argument against it, that tendency at least provides us with a good reason to try to find an alternative analysis of perception which doesn't have it.

The problem with this move is that there *is* no such alternative analysis. As Michael Lockwood has pointed out (1989, pp. 142-143), it is simply false to suggest that the skeptical problematic is uniquely threatening to an indirect realist construal of our perceptual contact with the external world. What gives rise to the skeptical problem is the fact that it is logically possible that my experiences could be just as they are now, when I take myself really to be seeing a computer, desk, lamp, window, etc., and yet I am not really seeing such things at all, but only hallucinating them, or dreaming them, or being caused by a Cartesian evil genius or a supercomputer to have experiences as of seeing them, and so forth. And this fact holds regardless of whether indirect realism or direct realism is true. Let our awareness of physical objects in veridical perception be as direct as you wish: it is still an open question whether, in any particular case where you *think* you're having a veridical perception, you *really* are, or can be *justified* in believing that you are. So the suggestion that indirect realism should be rejected, because it would lead us into a skeptical problematic, really cuts

little ice. That problematic is with us *whatever* position we take. It poses no *special* difficulty for the indirect realist.

Someone might think to retort: "But if we're *never* directly in contact with the external world, this might make it much harder to refute skepticism. For if we can know that it is at least possible directly to be aware of a physical object, and can know what it would be like so to be aware – as direct realism says we can – then it seems there is, on the direct realist position, at least more to work from in constructing a response to skepticism than there is on the indirect realist view that we can't know this."[7] I'm not sure that such a suggestion could actually be developed in a way that would make hay against the skeptic, but even if it could be, I think its force is at least cancelled out by an advantage indirect realism has over direct realism vis a vis skepticism, namely that the former view better accounts for the fact that there is a skeptical problem at all. For if we're never directly aware of anything but our own qualia, it is perfectly understandable why there should be occasions when we think there are external objects corresponding to those qualia when in fact there are not. The fact of, and nature of, hallucination and the like becomes intelligible. But if we are usually directly aware of external objects, it is puzzling why we should sometimes

[7] Another way to argue that indirect realism at least makes skepticism more difficult to refute would be to appeal to another view associated with McDowell to the effect that if we are never directly aware of the external world, that world would not only be unknowable, but *unthinkable*: we couldn't so much as form a conception of it. But Lockwood has, I think, shown that this isn't so (1989, pp. 300-301): Imagine the case of a person who has spent his entire life encased in a tight-fitting suit with tiny television monitors fitted into the goggles, tiny speakers built into the ear pieces, pressure plates built into the body of the suit which stimulate the skin, etc., so that his awareness of external objects

have experiences that are just like the veridical ones but in which we are not aware of any external objects at all, and it is puzzling why those non-veridical experiences should be so much like the veridical ones. Indirect realism, I suggest, thus has greater explanatory power than direct realism.

It is also not at all obvious that there *is*, as is often supposed, no way to justify belief in the external world against the skeptical challenge. One well-known way of trying to do so is to appeal to the hypothesis that there are external objects corresponding to our experiences as the *best explanation* of those experiences, one that is constantly confirmed given that the predictions we make on the basis of it generally pan out. As Lockwood argues (1989, p. 298), this sort of defense is exactly parallel to the scientist's justification of hypotheses about such unobservable entities as electrons. Thus, if belief in electrons is rationally justified by the fact that hypotheses positing them are well confirmed, despite the fact that electrons are not directly observable, then belief in external physical objects can be similarly justified, despite the fact that *they* are not directly observable.

One response to Lockwood's suggestion might be that it isn't clear that the common sense hypothesis of external physical objects really is the *best* explanation; for maybe the skeptic could argue that a Cartesian demon or evil genius explanation, say, has the same degree of explanatory power, but is simpler than the common sense view, and is thus to be preferred. After all, unlike the common sense view, which posits an

---

is always mediated by his awareness of what takes place within the suit. Such a person would never

enormous number and variety of external objects, governed by complicated laws, "the demon hypothesis proposes just one object (the demon) operating according to a principle (the desire and pursuit of deception) that we are intimately acquainted with... [so that it is] both simpler, and more intelligible, than the doctrine of common sense" (Scruton, 1994, p. 20). But the best response to this, I think, is to point out that the very same thing could be said against scientific hypotheses which posit electrons and the like. "Why accept a theory of the physical world which appeals to a myriad of unobservable particles, operating according to complicated laws of the sort physics tells us about, when we can opt instead for a more economical Cartesian demon explanation?" a skeptic about physics might say; "Maybe it's the demon who's responsible for the observable phenomena usually explained in terms of modern physics, chemistry, and so forth." If such an outlandish suggestion is to be rejected as a serious alternative to modern science, then it is hard to see why it shouldn't be rejected also as an alternative to the common sense view of what explains our experiences. But if a Cartesian demon explanation is thought to be a genuine rival to that common sense view, then it is no less a rival to modern science. We see again that the skeptical problematic doesn't threaten merely a particular kind of philosophical theory, nor is it relevant only to a particular domain of inquiry, despite the fact that philosophers tend to suppose otherwise.

---

be directly aware of the external world, but clearly he would still have a conception of it.

There thus seems little reason to doubt what there is in any case ample reason to believe, namely that we have no direct knowledge of the external physical world, and that what we do know of it is mediated by our direct awareness of our own qualia. But indirect realism is only part of the Russellian picture of our knowledge of the external world; or in any case, it has implications for that knowledge which go beyond its being merely indirect. "If causal [or indirect] realism is true," Lockwood writes, "the material world is, in a sense, *inscrutable*" (1989, p. 156). He elaborates as follows:

> Ultimately, we can know objects only by their sensory effects. Believing in external causes, we credit material objects with systematic causal powers to produce certain sorts of sensation in us. And the words 'red' or 'round' signify those intrinsic attributes, *whatever they may be*, which ground the powers or causal dispositions of an object to produce in us visual sensations of redness or visual and tactile sensations of roundness. In one clear sense, we do not, and indeed cannot, know what external objects are like in themselves. We know, or take ourselves to know, the causal structure of the world; but we do not know how this causal structure is qualitatively fleshed out. (1989, p. 155)

Slightly less opaquely, what we can know of the external physical world, on the Russellian view, are just the relations the objects within it bear to one another, relations described most precisely in the mathematical language of physics. Physical objects, events, and processes are just those objects, events, and processes, whatever

they are, that fill the nodes of the abstract causal structure in terms of which physics represents the physical world. An electron, for example, is described by physics entirely in terms of the role it plays in a physical system, i.e. its causal propensities, its relations to other particles, and so forth; but what *exactly* it is that plays this role is something physics doesn't tell us, nor is it something perception tells us. Indeed, we *cannot*, through perception or physics, know what it is: we cannot know the *intrinsic* nature of electrons or of any other physical object, event, or process – that is, we cannot know what it is like in itself, apart from its relations or abstract structure. For what I know in perception are just certain mental phenomena, qualia. I also believe – and if the sort of answer to skepticism described above succeeds, can know – that there are objects, events, and processes external to my mind which produce those qualia. But since I can't know them directly, all that is left for me to know about them is the postulated causal relationships they bear to my own qualia; and, derivatively, through scientific inquiry, their postulated relationships to one another.

We might call this the thesis of "the inscrutability of matter" or, better, given that it allows us at least knowledge of the physical world's causal structure, "the structuralist thesis" – to borrow some terminology from John Foster (see his 1982, Chapter 4, and 1991, p. 123, respectively), a philosopher who endorses it though he rejects the rest of the Russellian view. There are differences among Russellians over the precise degree to which our knowledge of the external physical world is limited to knowledge of structure. For example, Galen Strawson (forthcoming) would allow

that we have some knowledge of the intrinsic nature of physical *space*.[8] Suffice for now to say, however, that all Russellians are committed to some version or other of the structuralist thesis, and all agree that that thesis opens the way to a novel defense of the mind-brain identity theory – and along with it, a novel solution to the qualia problem.

(b) The Russellian mind-brain identity theory

Paul Churchland tidily sums up the standard case for the mind-brain identity theory in his Matter and Consciousness (1988, pp. 26-29). To all appearances, a human being, or any other animal thought to be conscious, begins its existence as a purely physical entity, i.e. a fertilized ovum, and develops from it by purely physical processes; and human beings and other animals are also products of the evolutionary process, which itself has purely physical beginnings and proceeds by biological principles operating in accordance with physical laws. Moreover, all mental phenomena appear systematically to be dependent on physical phenomena, and in particular on neural phenomena, such that alterations to the latter (such as brain injury, ingestion of drugs, and so forth) inevitably have an effect on the former; and neuroscience has made steady progress in explaining human behavior, mental illness,

[8] Perhaps *some* such qualification is needed in order to meet the objection, originally suggested by the mathematician M.H.A. Newman and cited by A.C. Grayling (1996, p. 63), that if our knowledge of the external world was *entirely* limited to structure, we wouldn't be able to differentiate one of any number of structurally isomorphic possibilities as the *real* world, so that we wouldn't in fact have the substantive knowledge of the physical world we take science to give us, and which even Russellianism takes us to have (albeit to a limited extent). In any case, what the Russellian view really needs, as we'll see, is not a *complete* lack of non-structural knowledge, but only enough of a lack to undermine any claim that mental qualities couldn't possibly be identical to physical qualities. Indeed, as we'll

and much else in neurological terms. In short, the evidence appears strong that human beings and other conscious creatures are purely material systems, so that consciousness itself must somehow be a purely material phenomenon – a phenomenon clearly tied specifically to the brain and central nervous system. But then, just as the links discovered between temperature and mean molecular kinetic energy led scientists to *identify* the two, or to *reduce* the one to the other, so that it turns out that warmth *just is* high average molecular kinetic energy and that coolness *just is* low average molecular kinetic energy (and similarly for other well-known reductions, e.g. light to electromagnetic waves, lightning to electrical discharges, genes to DNA, etc. etc.), we are also justified in making another identity or scientific reduction in the case of the mind and the brain. Mental states, we should conclude, *just are* states of the brain and central nervous system. *The mind just is the brain.*

To this case Michael Lockwood would add another consideration drawn from modern science (again following Russell, but developing the idea more fully than Russell did). Einstein's special theory of relativity tells us that any two events separated in time with respect to one frame of reference must be separated by space with respect to another frame of reference, and events which are spatially *separated* must be spatially *located* (1989, pp. 72-73). But then since *mental* events, as would be almost universally acknowledged, exist in *time*, it follows that they must exist in *space* as well. In short, "the theory of relativity so mixes up space and time as to

---

also see, the Russellian view *ultimately* holds that we *can* have knowledge of the intrinsic nature of

make it incoherent to suppose that mental events could be in time without being in space" (1989, p. 78). And this opens the way to the possibility, in principle at least, of giving a spatial location to mental events, if one knows their temporal location. All this tends to undermine a common objection to the identity theory that it is odd, and even meaningless, to assign a spatial location to mental events, at least independently of *presupposing* that they are identical to brain events (so that any argument for an identity of mental and physical events on the basis of identity of spatial location would be question begging). (Not that this objection was absolutely decisive in the first place: it does indeed sound at least slightly odd to assign a precise spatial location, e.g. "two inches behind the left eyeball," to one's thought that it is sunny outside – though most people, including most non-philosophers, probably wouldn't bat an eye at at least the vague suggestion that one's thoughts are "in one's head" – but it is, of course, perfectly natural to say that one's pain is in one's back, or that the sensation of sweetness one has on tasting an apple is in one's mouth, and so forth.) And given the very general (temporal) correlations between brain states and mental states, it also tends to provide some positive support for an identity of mind and brain.

Of course, to this suggestion, as to the considerations summarized by Churchland, someone like Searle might respond that the most we have *reason* to suppose is that the mind is *dependent* on the brain, but not that they are *identical*, and that given the objections to physicalism we've already looked at, this is in any case the

---

the material world, only not through perception or physics.

most we *should* suppose. But as we've seen, Searle's allegedly equally naturalistic alternative to physicalism is hardly without problems itself: it's not as if we have evidence that supports either the problematic identity theory or Searle's biological naturalism, and that the latter "wins" by virtue of lacking difficulties of its own. Nor is it clear how Searle's view could accommodate the considerations Lockwood brings to our attention. At least on the identity theory, we have a clear-cut answer to the question of where various mental phenomena are located: they're in the brain, specific mental states having more-or-less specific locations in the brain (even if some of these are best thought of as spread out over the brain, as widely separated processes might underlie a particular mental state). But on Searle's view, on which mental states are higher-level features of the brain which are nevertheless not identical with any of the neural processes taking place within it, it's much less clear what to say. Is my belief that it's sunny outside, if not identical with, and thus locatable at the same point as, a specific brain state, to be located somewhere else close by, say, floating somewhere an inch or two *above* the portion of the brain in which the correlated brain state exists? Presumably not!

Nevertheless, as we've seen, despite the powerful case to be made for an identification of the mind with the brain and the arguably insurmountable difficulties with alternative views, there are also seemingly equally insurmountable difficulties with such an identification, as with all other forms of physicalism. But now at last we come to the Russellian suggestion for breaking this deadlock, a suggestion which appeals to

what we've concluded regarding the nature of our knowledge of the external physical world to defend a *non-physicalistic* version of the identity theory. Call it the *Russellian identity theory* – or the RIT for short.

The RIT, like other identity theories, holds that mental states are identical with states of the brain. But, in a sense, it turns standard identity theories on their heads. As we've seen, the standard criticism of identity theories is that there is a gap between the facts about the brain and facts about the mind, particularly facts about qualia, which is such that it seems possible that all the former facts could be just as they are, and yet none of the latter facts obtain. Subjective, first-person facts about the mind seem to be facts over and above the objective, third-person facts about the brain. Identity theorists typically try to respond to this by arguing – unsuccessfully, I have tried to show – that mental properties can somehow be reduced to physical ones, that apparently irreducibly first-person features can be shown somehow to be nothing but third-person features (or eliminated altogether). But on the RIT, given its commitment to indirect realism and the structuralist thesis, we have no knowledge of the intrinsic qualities of the third-person realm of external material objects to begin with; and it follows that we lack any basis whatsoever for holding that the third-person facts could be just as they are, and yet the first-person facts could fail to obtain. *We simply lack the knowledge requisite to being able with any confidence to make such a judgement.* What we *do* have knowledge of is the intrinsic nature of those first-person facts themselves – of our own mental states, including our qualia. And since these

mental states are, ex hypothesi, identical with states of the brain, it follows that we have knowledge of the intrinsic nature of at least one material object, the brain (not a material object *external* to the mind, however, for the obvious reason that, on this view, it *is* the mind). Our awareness of qualia, then, is really awareness of the intrinsic qualities of the brain, which are produced ultimately by external material objects of whose intrinsic qualities we can have no direct knowledge. So whereas on the typical identity theorist's view, matter is unproblematic, and mind somehow has to be accommodated to it, on the RIT, it is mind, namely the mental properties of the brain, that are directly known, and the material world external to it is, as Foster puts it, "inscrutable." As Russell himself so colorfully (if outrageously) summed up this theory:

> I should say that what the physiologist sees when he looks at a brain is part of his own brain, not part of the brain he is examining. (1954, p. 383)

This statement, of course, needs certain qualifications. What Russell means is that the physiologist is not *directly aware* of the brain he is examining, though of course he is aware of it *indirectly*; what he *is* directly aware of are qualia – which are identical with states of his own brain.

Lockwood (1989, p. 160) notes that the RIT in effect inverts a strategy physicalist identity theorists like Smart used to defend their version. Smart (1959) famously argued that mental ascriptions are *topic-neutral*, by which he meant that when we ascribe a mental state, the having of a particular quale, say, to someone, we

are ascribing to him a state the accurate description of which makes no reference to the intrinsic nature of the state, but only to a particular role it plays. So to say that someone is having a yellowish-orange after-image is just to say that there is something going on in him which is like what goes on when he's looking at an orange in good light, and so forth. And this leaves it open that what answers to this description is, as the identity theorist holds, in fact a neural state or process. It turns out, on the physicalist identity theory, that brain states are what play the roles we describe (such theorists claim) in a topic-neutral way when we talk about mental states. The mind turns out to be the brain. On the RIT, however, it isn't our conception of *mental* states which is topic-neutral, but rather our conception of *physical* states, including *brain* states. What we know about the brain, in knowing what we know through perception and neuroscientific research, is not its intrinsic nature, but only is abstract structure: we know that it is a complex system of events having certain causal relationships to each other and to our perceptions, a network of states which play certain *roles*. But what it is exactly which *plays* those roles we don't know – at least not *directly* through perception. We do know the intrinsic nature of our mental states, however; and we know that there are general correlations between those states and states of the brain. And this provides us with a justification for supposing an identity between the two. That is, it provides us with justification for concluding that it is mental states – qualia and the like – which play the roles abstractly or topic-neutrally described for us by neuroscience and known to us through perception. So mind and

brain are identical after all: but it's not that mental states turn out to be (or are "reduced to") brain states so much as that *brain* states turn out to be *mental* states! The brain turns out to be the mind.

The RIT, no less than the sort of physicalism represented by Churchland, thus insists that the introspection of qualia is "the direct introspection of brain states." But it gives a radically different interpretation to this phrase, one which would no doubt make Churchland and like-minded philosophers cringe. For in making qualia – construed not in a reductionistic way but in all their subjective, first-person, qualitative glory – out to be the stuff of the brain, the RIT is a mind-brain identity theory even a dualist could love, or at least stand the company of. And it is *non*-physicalistic precisely because it does so. It denies that the third-person facts revealed by physics, even a completed physics, are all the facts there are.[9] Indeed, it insists, far from that being the case, what physics does reveal to us about the world is surprisingly very little. Moreover, the naïve conception of matter shared by common sense and most philosophy of mind alike is entirely misguided. We (quite understandably) tend to think of matter pre-reflectively as the sort of thing we are familiar with in everyday

[9] This more or less follows Lockwood's conception of physicalism, on which it is the view that all the facts about the mind that there are are the sorts of facts with which physics deals (1989, p. 18); though on Lockwood's conception of *materialism*, that view merely "denies that mental states, processes, and events exist *over and above* bodily states, processes, and events" (p. 20), so that he is quite happy to call himself a (non-physicalistic) materialist. But Maxwell, whose view is essentially identical with Lockwood's, reverses the senses of these terms, describing his position as "nonmaterialistic physicalism" (1978, p. 365)! Because of these inconsistencies, and because "materialism" and "physicalism" are often used interchangeably and in any case to label positions which are radically at odds with the Russellian view, I have not tried to salvage either as an appropriate label for the view. (Incidentally, to make matters worse, Chalmers, who at least flirts

experience – as the bulky, colorful, hot or cold, heavy or light, fragrant or malodorous, sweet or sour stuff we imagine whenever we think of paradigm cases of physical objects. We think of brains, in particular, as the gray, wet, squishy globs we've seen photos of, or perhaps even had the chance to see and touch up close. But in fact, none of these qualities we imagine when we ordinarily conceive of the material world is a quality of material objects as they are in themselves (to speak in Kantian terms) – at least, we have no basis for thinking otherwise. Such qualities are instead the qualia with which we are presented when external material objects affect our senses; and they are themselves *features of our own brains*. When we imagine the qualities by which we typically characterize external physical objects, then, we are really imaging qualities of our own brains, features of the internal world which are regularly caused by those external objects. What we really know about the external physical world, including brains considered "from the outside," as physical objects alongside other (external) physical objects, is merely its abstract structure. The *intrinsic* nature of the material world, or at least that part of it which constitutes our brains, is known to us, not through physics, but through *intropsection*. The common sense intuition, which dualists capitalize on and physicalists try unsuccessfully to explain away, that physical objects, events, and processes have properties which are incompatible with those had by thoughts and sensations is thus unfounded.

---

with Russellianism as a possible way to flesh out the precise relationship between mind and matter (1996, pp. 303-308), is a *property dualist*!)

The Russellian response to the objections made to physicalist versions of the mind-brain identity theory should be clear. The claim of the knowledge argument, it will be recalled, was that one could have a complete account of the neurophysiological facts concerning human beings or animals and still not know what it's like to experience red, say, or what it's like to have the experiences bats have in getting about the world by means of echolocation, so that there must be something more to the mind than just the brain. But if all this neurophysiological knowledge amounts to in the first place is just knowledge of the causal structure of the brain, then it just isn't all there is to know about the brain after all. In particular, it isn't knowledge of what sorts of things fills the nodes of this causal structure or play the causal roles the neurophysiology tells us about. And on the RIT, it is precisely qualia which do so, precisely the elements the knowledge argument says a mind-brain identity theory must leave out.

A similar reply can obviously be made to the zombie argument. For it turns out that to imagine beings identical to us neurophysiologically, behaviorally, and so forth, is just to imagine a certain sort of causal structure. It is not to imagine what sorts of things play the causal roles within that structure. And so it does not, by itself, amount to imagining beings whose brains are just like ours. Making it amount to this, on the RIT, would involve imagining further that what play the causal roles in question are qualia. But then we wouldn't be imagining "zombies" at all, and thus we wouldn't be imagining anything that could serve as a counterexample to the theory. Arguments

like Kripke's can be rebutted in like fashion. His argument alleges that, since expressions referring to mental events and expressions referring to brain events are rigid designators, any statement of an identity between a mental event and brain event, if true at all, would have to be a necessary truth. But since, it is alleged, it is possible that a given brain event could occur without the corresponding mental event occurring (e.g. it is possible that, say, C-fiber stimulation, or whatever, is occurring but pain is not), such an identity statement couldn't be necessarily true. So all such statements must be false. But this argument depends, of course, on the supposition that we know enough about the intrinsic nature of brain events to be able to judge that they could be occurring without mental events also occurring; and this, as we have seen, is just what the RIT denies. On that view, all we know about brain events, considered as material events, is just what we know about all material events, i.e. their causal interrelationships. So what it is to be a certain type of brain event is just to play a certain causal role relative to other material events. And given this characterization of brain events, an identity between brain events and mental events (say between C-fiber stimulation, understood abstractly as whatever type of event plays such-and-such a causal role, and pain as the type of event that turns out to play it) is no more problematic than any other identity claim made in science (such as the claim that, to put it crudely, genes are DNA, that is, that the causal role specified by talk of genes turns out to be played by DNA).[10]

[10] Maxwell (1978) provides a much more detailed reply to Kripke from a Russellian perspective.

Its radical rejection of the standard conception of the material world thus allows the RIT to sidestep the difficulties that seem fatal to physicalism in all its forms, while at the same time, in identifying mind and brain, staying within the bounds of a broadly naturalistic framework. But just here, it might be objected, is the rub: the "naturalism" that results seems just *too* broad. So far out of its way does the RIT go to accommodate the dualist's scruples about qualia, it might be alleged, that it seems to do away with the *physical*, or at least to redefine *it* in terms of the mental, putting it at precisely the opposite extreme from that of eliminative, or at least reductive, materialism. For by making qualia out to be the intrinsic properties of the brain, the way is opened to the conclusion that they are the intrinsic properties of all *other* material objects as well, i.e. the ones external to the mind and known to us only indirectly. "Matter" arguably turns out, on the RIT, to be quite literally "the stuff of which dreams are made"! – in which case what we're left with would hardly seem to be a variety of naturalism at all, but rather a kind of *idealism* or *panpsychism* on which ultimate reality is mental through and through.

Lockwood is the Russellian thinker who has dealt with this objection in the greatest depth, and he is particularly sensitive to it because he endorses the notion that our introspection of qualia gives us a grasp, indeed our only possible grasp, on what the intrinsic nature of the material world external to the mind is like. But he denies that the objection is fatal, on the basis of a position that is as bold and radical as the RIT itself. Qualia, he argues, can exist independently of any subject, unsensed by any

perceiver who is aware of them or aware of external objects through them.[11] And if qualia can exist whether or not they are sensed by some perceiver, they cannot be thought of as *essentially* mental. Consequently, a world of objects the intrinsic qualities of which are qualia would not thereby be a world that was entirely mental in its constitution. The slippery slope from Russellianism to idealism or panpsychism would thus be blocked.

Lockwood's position here is another one which he adopts from, but takes much farther than, Russell himself; and it is that feature of the Russellian view which has led many commentators to give it a distinct classification all its own, rather than making it out to be a variety of identity theory or naturalism. For if qualia can exist unsensed by any perceiver, the way is open to a general metaphysical position that stands between materialism and idealism without being dualistic. Materialism holds that matter alone exists, and that putative mental phenomena are just a special class of material phenomena. Idealism reverses this, holding that mind alone exists, and that so-called material objects are in fact reducible to mental phenomena. Dualism, of course, takes mind and matter to be equally fundamental. But if qualia can exist unsensed, they can be seen as essentially neither mental nor material, but rather as

[11] I should note, however, that because of the usual connotation of the expression "qualia" as "the way things appear," as in "appear to a conscious subject," Lockwood prefers not to use this term, and speaks instead of "phenomenal qualities." But since "qualia" has come to be so commonly used (beating out not only "phenomenal qualities" but also "sensory qualities," "sense-data," "sensibilia," "sensa," and God knows how many other expressions with identical or at least similar, overlapping connotations), I'm going to stick with it, and ask the reader to keep in mind that Lockwood's use of the term would not commit him to any assumption that qualia can only exist in the minds of perceivers.

falling under a more basic category. Qualia, on the view Lockwood defends, are the intrinsic qualities of the objects that make up the world, both those that are associated with awareness (minds, or brains) and those that are not (ordinary material objects). That view may thus be regarded as a species of *neutral monism*, a metaphysical position according to which (contra dualism) there is only one basic kind of "stuff," which is (contra materialism and idealism) neither mind nor matter, but something neutral between them, and out of which both mind and matter are constructed.

As I've said, this conception of qualia is one that Russell himself also held, at least at one point in his career; and Russell was indeed also at that point a proponent of neutral monism. There is, however, some controversy over how closely he did, or would need to, associate these views with the version of the identity theory we've been examining. Lockwood has argued at length (in his 1981, an interesting and important study of Russell's development which predates Lockwood's own vigorous advocacy of the positions discussed therein) that the position Russell took post-1927, namely what we've been calling the RIT, is more or less continuous with the neutral monism he advocated in such earlier works as The Analysis of Mind (1978), albeit that there were some important developments in between – contrary to the view of most Russell commentators at least up to the time Lockwood wrote.

Russell's earlier neutral monism, as I understand it, might be characterized as follows. What we are directly acquainted with in perception are sense-data, or, again, more or less what would generally now be referred to as "qualia"(to ignore some

subtleties). But these sense-data or qualia are not (as for indirect realism) caused by external physical objects, for there *are* no objects external to them. Rather, physical objects just are *collections* of sense-data; they are, that is, "logical constructions" out of sense-data. So far this sounds like phenomenalism or Berkeley's idealism. But unlike these doctrines, neutral monism holds that sense-data are not essentially mental, for they can exist apart from any subject, and indeed, there *is* no subject existing over and above them, to which they must be present in order to exist. Minds too, no less than physical objects, are logical constructions out of sense-data. So there is only one kind of stuff (hence the label "monism") which is however neither mental nor physical (hence the adjective "neutral").[12]

What changed between the time Russell took this view and the time The Analysis of Matter appeared in 1927 was that he came to adopt scientific realism and the indirect realist theory of perception, so that he came to see physical objects as entities which do, after all, exist apart from, over and above, the qualia with which we are acquainted in introspection. Moreover, he now identified these qualia with states of the brain. Because he nevertheless did (or could) see even external physical objects as having unsensed sense-data or qualia as their intrinsic properties, so that the latter are still not to be thought of as exclusively mental, Lockwood classifies Russell's identity theory as also a kind of neutral monism.

---

[12] Hence it is in my view misleading – though understandable given the oddness of the notion of unsensed sense-data – to classify neutral monism, as some have done, as merely a "notational variant on idealism" (Snowden, 1995).

There is much to be said for this interpretation, and reason to think Russell himself did think of his later view as a variety of neutral monism, as any reader of Lockwood's 1981 knows. Still, there is also reason to think "neutral monism" is, historical usage aside, not an appropriate label for the RIT. For unlike Russell's earlier neutral monism, the RIT doesn't hold that there is some neutral stuff more basic than the physical world, out of which both it and mind are constructed. Rather, it, no less than physicalism, takes the physical world itself to be basic, and mind to be just a special kind of physical thing – though of course, it radically reconceptualizes the intrinsic nature of the physical.[13] More importantly for the issue at hand, it's not entirely clear that Russell himself maintained his belief in unsensed sense-data after adopting what we've been calling the RIT. (A.C. Grayling (1996, pp. 60-61) is one interpreter who thinks he did not, and even Lockwood (1989, p. 160) grants that it is unclear how Russell conceived of the intrinsic nature of physical objects outside the brain.) In any case, it isn't obvious that there's anything in the RIT which *requires* this: Russell could just say that in knowing qualia, we know the intrinsic qualities of *one* physical object, namely the brain, but know nothing about any other physical object's intrinsic nature.

There would, however, be something unsatisfactory about this. If we hold that the brain, however uniquely complex, is nonetheless but one physical system among

---

[13] Perhaps for this reason, even Lockwood seems to have stopped using "neutral monism" as a label for the Russellian view in his later writings on the subject (1989, 1993, 1998), though to my knowledge, he has never explicitly renounced the label.

others, and want also to fit qualia into the natural world in a way that removes the air of mystery surrounding them – the sense that they are somehow fundamentally different from anything else that exists – then we have grounds, and motivation, for ascribing the intrinsic qualities of the brain to other physical systems as well. And if we want to do so and avoid panpsychism or idealism at the same time, it seems we are bound to accept the notion of unsensed qualia.

Whatever view one takes about the terminological issue of whether to call the RIT a version of neutral monism, then, I think Lockwood is right on the *substantive* claim that the doctrine of unsensed sense-data or qualia is an essential part of the RIT – at least if that theory is to avoid the slippery slope to panpsychism.

So much, then, for setting out the Russellian solution to the qualia problem. What remains to be seen is whether or not it is *true*. And the first order of business is determining whether or not the doctrine of unsensed sense-data is defensible – and if it isn't, whether or not the resulting slide into panpsychism itself constitutes a fatal flaw.

5. Troubles with Russellianism

(a) Could there be unsensed qualia?[1]

As we've seen, Lockwood has claimed that in order to avoid a slide into panpsychism or idealism, proponents of the RIT must commit themselves to the doctrine of unsensed sense-data or qualia, qualia which exist outside the awareness of any perceiving conscious subject. It is now time to examine his arguments for this doctrine. In raising the issue, it should be noted, Lockwood has actually reopened an almost entirely (and unjustly) forgotten debate carried on by the sense-datum theorists of the early twentieth century, among whom are to be counted not only Russell, but also such thinkers as A.J. Ayer, C.I. Lewis, G.E. Moore, and H.H. Price.

First, Lockwood's own precise characterization of the thesis should be noted. He calls it the "disclosure view," which he says "might be thought of as a kind of *naïve realism* with respect to phenomenal qualities" (1989, p. 162). That is, phenomenal qualities or qualia are intrinsic attributes of states of the brain, which exist independently of their being sensed, and which awareness now and then discloses to us. They are not qualities *of* awareness itself. Says Lockwood:

> On this view, phenomenal qualities are neither realized by being sensed nor sensed by being realized. They are just realized, and sensed or not as the case may be. The realization of a phenomenal quality is one thing, I contend; its

[1] Much of the material in this section has appeared earlier in my 1998a, and has benefited from helpful comments on the original paper made by two anonymous referees.

being an object of awareness is something else, albeit something for which its realization is a necessary condition. (1989, p. 163)

Now some might want to cut the debate short at this very point, before any arguments in favor of the existence of unsensed qualia even get going, on the grounds that the very notion is a confused one. Lockwood cites A.J. Ayer and C.I. Lewis as two philosophers who take such a view (1989, pp. 170-1). Another seems to be D.M. Armstrong, who argues that the way the notion of qualia (he uses the terms "sense impressions" and "sense data") is introduced rules out their existing unsensed (1961, pp. 35-7). For the notion of a quale, he says, is just the notion of the way something seems to be to a person in perceptual experience, whether the experience is veridical or illusory. How, he asks rhetorically, could such a thing exist apart from a mind that has it?[2] This sort of consideration is not without force. For it does seem to be true that the notion of a quale is often developed by reference to presumably purely subjective mental phenomena, such as illusions, hallucinations, inverted spectra, and the like. Other terms used by philosophers for these qualities, e.g. "raw feels," "sensations," etc., reflect this tendency. Lockwood says that "the ontological status of these qualities seems to me...to be a substantial matter of fact; not something to be decided...simply by linguistic fiat" (1989, p. 171); and it is no doubt good philosophical practice to be wary of any attempt quickly to settle a philosophical

[2] As Galen Strawson notes (1994, p. 129), Gottlob Frege held a similar view, namely that "an experience is impossible without an experiencer" (Frege 1967, p. 27). Accordingly, Strawson labels this position "Frege's Thesis."

dispute by a facile appeal to the way words are used. Still, not all such appeals are facile, and when terms are used in a way that appears paradoxical given the way they are normally introduced, it seems reasonable to put the burden of proof on the innovator to show that the appearance of paradox is misleading.

Lockwood does say one other thing about this sort of objection (1989, pp. 164-5). He cites Berkeley's notorious argument to the effect that since it is impossible to abstract the idea of the tree that one perceives from the idea of one's perceiving it, for the tree to exist is for it to be perceived. And he says that if there is no good reason to think that this abstraction is impossible, neither is there any reason to think it impossible to abstract the idea of a phenomenal quality from the idea of one's being aware of it. If Berkeley's argument is bad, so is that according to which the notion of unsensed qualia is just confused. Lockwood admits of "one salient disanalogy" here, namely that though we can conceive of a tree without conceiving what it would be like to see one, we cannot conceive of e.g. phenomenal red without conceiving what it would be like to be aware of it; but he says that this "merely reflects a difference in the nature of what one is required to grasp in each case" (1989, p. 165). We may well wonder here, though, whether Lockwood has thus given up the game. For what he has just described seems to be the notion of something that (like a tree) exists apart from any mind, but (unlike a tree) cannot be conceived of except as being present to a mind. We may, I think, be forgiven for at least doubting the coherence of such a notion.

All in all, then, the preliminary linguistic or conceptual considerations, even if we grant that they are not decisive, appear to put the burden of proof on anyone who wants to claim that qualia can exist unsensed. Such a theorist cannot simply assume a parity between his position and that of his opponent, and proceed in the assurance that the coherence, if not the truth, of his position can be taken for granted. A positive case must be made for unsensed qualia if we are even to consider them a live possibility. Lockwood does try to develop such a positive case, and it consists, as far as I can tell, of two arguments.

The first argument goes as follows (1989, pp. 163-4). Consider three patches of color, projected on to a screen, labeled L (left), M (middle), and R (right), and suppose that they are such that in the absence of R, L is indistinguishable from M, and in the absence of L, M is indistinguishable from R, but L is always distinguishable from R. Now it can't be, Lockwood says, that the corresponding phenomenal patches of color (the qualia one is aware of in the experience of looking at the patches on the screen) are such that the left and middle ones are qualitatively identical, the middle and right ones are as well, but the left and right ones are qualitatively distinct. So how are we to describe what happens phenomenally when one is looking at a screen that first contains only L and M, then L, M and R, and finally only M and R? The most plausible description, Lockwood says, is one in which:

> [T]he phenomenal colours corresponding to L and M are distinct, even in the absence of R: there *is* a phenomenal difference here, but one which is too

> small to register in consciousness, no matter how closely the subject attends. Adding together two phenomenal differences of this magnitude does, however, produce a difference that registers in consciousness; hence the subject's ability to distinguish L from R. (1989, p. 164)

We must conclude, then, that the characteristics of the qualia one is aware of can outrun one's awareness of them. That is, qualia can have attributes that the one aware of them is not aware of their having. And this, Lockwood thinks, gives us reason to think that qualia may "quite generally...outrun awareness," that is, that they may exist unsensed by any perceiver (1989, p. 164). Furthermore, qualia that exist in, say, portions of one's visual field that one is not conscious of, due to inattention, provide a model for such unsensed qualia, and thus for what the unsensed portion of the physical world is like in itself, even beyond the brain.

This is all, I think, much too quick. To be sure, Lockwood's example does indeed appear to show that qualia can have attributes of which we are not aware. For since the phenomenal patches of color corresponding to L and M differ in respect of shade even when this difference is undetected, each has an attribute of which a conscious subject aware of these qualia themselves may not be aware: The phenomenal patch of color corresponding to L, for example, has the attribute of "differing in respect of shade of color from the phenomenal patch of color corresponding to M." But it does not follow from this that there are qualia (as opposed to mere attributes of qualia) of which no one is aware.

That conclusion would follow only given a further thesis, namely that attributes of qualia are themselves qualia. But are they? The fact that Lockwood argues as he does is some indication that he thinks so. But he doesn't explicitly state this thesis, much less argue for it; and there appears to be no reason to think that it is true. Certainly, the example under consideration gives it no support. "Differing in respect of shade of color from the phenomenal patch of color corresponding to M" is not a quale.[3]

Furthermore, even if it could be shown that there are some attributes of qualia which are themselves qualia, this would not be enough to help Lockwood. For it might be that, though it is possible to be unaware of some attributes of qualia that one is sensing, there are other attributes of which one cannot fail to be aware, namely those attributes which are themselves qualia. At the very least, then, what is needed for Lockwood's argument to go through is an example of an attribute of a quale which is both a quale itself and unsensed by a conscious subject who was aware of the quale of which it is an attribute.

But in fact, it is arguable that even this sort of example wouldn't do the job, at least not well enough for Lockwood's ultimate purposes. For suppose there are such examples. We can then distinguish between first-order qualia (such as the phenomenal

[3] G.E. Moore held explicitly that attributes of qualia or sense-data are *not* themselves qualia: "I should now make, and have for many years made, a sharp distinction between what I have called the [phenomenal] 'patch' [of color], on the one hand, and the colour, size, and shape, *of* which it is, on the other; and should call, and have called, *only* the patch, *not* its colour, size or shape, a 'sense-datum' (1962, p. 44, n. 2. This note is a gloss on an earlier text in which Moore seems to imply the

patch of color corresponding to L), and second-order qualia (attributes of first-order qualia which are themselves qualia). The most that such examples would show is that there are second-order qualia of which no one is aware. But whether one is aware of them or not, second-order qualia can exist only when first-order qualia do, since they are attributes of the latter; and for all that the hypothetical examples would show, it may yet be that first-order qualia cannot exist unsensed, so that even second-order qualia would ultimately depend on some conscious subject for their existence. In short, even if there are attributes of qualia which are themselves qualia and which are unsensed , it wouldn't follow that they are not ultimately dependent for their existence on some conscious subject. But surely, only if qualia are not so dependent can they serve Lockwood's purpose of being the intrinsic qualities of a world that is not panpsychist, not mental through and through.

These considerations are supported by what was said earlier concerning Lockwood's defense of the very coherence of the notion of an unsensed quale. If, as Lockwood seems to grant, one cannot conceive of phenomenal red without conceiving of what it would be like to be aware of it, then surely one cannot conceive of phenomenal red's having attributes that he is unaware of without conceiving of what it would be like to be aware of phenomenal red and yet unaware of some of its

---

opposite view). And Moore was himself a philosopher who thought it at least possible that sense-data could exist unsensed (p. 58)!

attributes. As I suggested above, it is difficult to see how something could be both conceivable only as an object of awareness and exist independently of any mind.[4]

All things considered, then, the most that Lockwood's opponent need grant in the face of the argument under consideration is that qualia, which do not exist unsensed, can have attributes one is not aware of their having.

Furthermore, Lockwood's claim that portions of a visual field to which one is paying no attention provide a model for unsensed qualia is dubious. For one thing, a visual field is, of course, the visual field *of* some conscious subject. In the case of someone who is unconscious, we do not say that he is unconscious of what is in his visual field, but rather that there is nothing in his visual field, in fact that he has no visual field, at that time, at all. Surely, then, the most plausible reading of the kind of case Lockwood has in mind is not that a person is unconscious of what is going on in part of his visual field, but that he is only (in Lockwood's own phrase) "dimly conscious" of it (1989, pp. 163, 166). But even if only dimly conscious of it, he is still

[4] An anonymous referee has objected that this argument "seems to conflate 'what it would be like' to be aware of a phenomenal quality with actually being aware of it." I'm not sure I understand the objection. I certainly do not mean to conflate *conceiving* of what it would be like to be aware of a phenomenal quality or quale with actually being aware of it. That would be as implausible as conflating conceiving of what it would be like to see a cow with actually seeing one. If the charge is rather that I conflate conceiving of what it would be like to be aware of a quale with conceiving of actually being aware of it, then I plead guilty, or at least no contest. But that conflation appears entirely innocent. Think of what it would be like to have a reddish afterimage. Now think of actually having a reddish afterimage. Haven't you just thought of the same thing twice? My criticism of Lockwood, then, is just that if one cannot conceive of a quale without conceiving of what it would be like to be aware of it, then one cannot conceive of it without conceiving of actually being aware of it. Therefore, one cannot conceive of a quale except as actually being present to a conscious subject. So one cannot conceive of an unsensed quale.

conscious of it, so that such cases do not lend any credence to the claim that there are qualia of which no one is conscious.[5]

Finally, it seems that qualia are inherently perspectival. The red image I have when I look at a tomato from the front is different from the one I have when I look at it from above. The position of my body relative to the tomato always determines a corresponding difference in the qualia I am aware of when I look at it. So the character of a given quale appears to depend on the point of view of the subject who is aware of it. It is thus difficult, if not impossible, to conceive of a quale that did not have this perspectival character, and that, accordingly, was not an object of awareness for some subject. In particular, it is therefore difficult to see how even portions of one's visual field to which one is inattentive can provide a model for parts of the material world from which awareness is absent.

Lockwood's second argument is more promising, but I think that it, too, ultimately fails (1989, pp. 165-7). Any explanation of behavior in terms of

---

[5] Some readers may object to Lockwood's expression "dimly conscious," considering it too imprecise to support the points I wish to make by using it, both here and later in the paper (as an anonymous referee apparently does, referring to it as "a bit amateurish"). But surely what is meant by it is clear. While I write this, I am conscious of a number of things in the sense that they are in my field of vision, and unconscious of other things in the sense that they are not. Of those things that I am conscious of, some are at the focus of my attention, e.g. the screen of my word processor, and some are not, e.g. the fan to my right. For I have been (intermittently at least) thinking about the screen as I write, but not thinking at all about the fan until now. But even though I was not thinking about the fan, I was still conscious of it in a way that I was not conscious of, say, the bookshelf behind me. (One indication of this is the fact that if I were asked whether the fan was on, it is at least possible that I could have replied "yes," while if I were asked whether there was anything unusual on the shelf, I would have to have said "I have no idea.") So though the fan is not at the center of my attention, neither am I unconscious of it in the sense in which I am unconscious of the bookshelf. What I am is, in Lockwood's phrase, "dimly conscious" of it. That seems to me an adequate enough characterization to make my point, namely that the examples of phenomenal qualities Lockwood is

subconscious mental states and events, he says, must be of the same general form as those put in terms of conscious mental states and events. If talk of subconscious mental states is not merely metaphorical, then such states must cause behavior in the same way conscious mental states do. But if so, then we must acknowledge that, like some conscious mental states, some subconscious mental states are associated with qualia; for there are, of course, conscious mental states whose associated qualia are causally efficacious. (Consider a soldier's experience of seeing a white flag, where the phenomenal patch of whiteness plays a role in causing him to refrain from shooting the enemy soldier carrying the flag in a way a phenomenal patch of some other color would not.) And if so, then there are qualia of which no one is conscious.

This argument will only work if there are indeed subconscious mental states whose causal efficacy with regard to behavior parallels that of some conscious counterpart itself associated with qualia. Obviously, not *all* conscious mental states involving qualia have subconscious counterparts: There are, for instance, no subconscious states we would call cases of *seeing*. (Blindsight cases, even if they are to be classified as literal cases of seeing, are no help to Lockwood here, for what makes them "blindsight" cases in the first place is that there are no qualia involved.) But are there any cases of conscious mental states involving qualia that play a role in causing behavior which clearly have subconscious counterparts that also cause behavior?

---

speaking of still depend for their existence on their being in the visual field of some conscious subject,

The alleged examples of such given by Lockwood are not convincing. One involves the familiar case of a motorist who is engrossed in conversation with a passenger and thus not paying attention to his driving or to the road, but is still doing a decent job of getting where he wants to go, avoiding collisions, changing lanes, and so forth. (He's on "automatic pilot," we might say.) Now when such a driver *is* fully aware of what is going on, paying strict attention to the road and his driving, we surely want to say that the qualia he is aware of when he looks at the road, and so forth, are causally efficacious in determining his behavior. (For example, it matters, when he looks at the traffic light, whether it is a phenomenal patch of red or one of green that he is aware of. The former will play a role in causing him to stop, the latter in causing him to continue on.) But then we must also say that there are analogous qualia involved in the case just described, when the driver is not paying attention. And these, Lockwood argues, would then have to be qualia of which he is not aware.

I don't find this at all convincing. In both cases, the same sorts of thing are in the driver's field of vision. And in both cases (as can be seen from the fact that he has a field of vision in the first place) he is conscious. So the obvious response open to a critic of Lockwood, as he himself points out, is just to say that the driver is only "dimly conscious" of some elements in his visual field in the one case, and fully conscious of them in the other (1989, p. 166). In either case, he is still conscious, however fully or dimly, so there is no question of there being qualia that no one is

---

and are thus dubiously spoken of as unsensed.

conscious of. Lockwood's only reply to this is to say that it "seems...just empirically mistaken; there is, as a matter of brute fact, no need for that to be the case" (p. 166). But this is weak, and gives no reason for preferring Lockwood's account to that of his critic. Nor does the critic's response amount merely to answering one unsupported assertion with another, equally plausible or implausible, assertion. For as I have already argued, the burden of proof is on Lockwood in the first place, given the dubious legitimacy of the notion of an unsensed quale. So we do, in fact, have ample reason to reject Lockwood's account in favor of the account he puts into the mouth of his critic: The latter accounts for the phenomenon Lockwood cites perfectly well without appealing to unsensed qualia, which we have no independent reason to think exist, and the very notion of which is dubiously coherent. In short, Lockwood's case is undermined by Occam's razor.

Lockwood's only other example fares no better. The qualia associated with anger, he says, surely play a role in causing any behavior resulting from the anger when one is consciously angry at someone. But the same must then be true of repressed, and therefore subconscious, anger. So there must in this latter case be qualia that the subject is unaware of. The problem with this is that it is not at all obvious that anyone ever acts out of anger that is completely subconscious. There are, of course, people who are angry at someone, but who will nevertheless not admit to anyone, not even to themselves, that they are. Such people are said to be "in denial." But it is usually clear from the behavior of such a person that he is angry: He will tense

up when the person he is angry at comes in the room, instantly get into a bad mood when he is mentioned, and so forth. And in these cases, it is clear that there are qualia experienced by the person, namely those that partly characterize the emotional state of anger. And even if the anger is completely repressed (if this is possible), it would seem that it does not in that case cause any behavior. It only does so when it "seeps out" at times like the ones mentioned. And at those times, it is also clearly associated with at least dimly conscious qualia. This case no more needs to be accounted for in terms of unsensed qualia than the motorist case does.

One thing more should be said about Lockwood's argument concerning subconscious mental states. If it was successful at all, it would succeed in showing only that there are unsensed qualia that somehow make up parts of the brain not associated with states of awareness, since on the RIT, even unconscious mental states would still be identified with states of the brain. It would not tell us anything about what the material world external to the mind/brain is like. The same might also be true of Lockwood's first argument, for that matter, considering what I said above about the perspectival character of qualia. For if qualia are inherently perspectival, then even if they could exist unsensed, they would still seem necessarily to be associated with some mind (and thus, on the RIT, some brain). So even if Lockwood could show that qualia could exist unsensed, it isn't clear that this would enable him to halt the slide toward panpsychism that he thinks the RIT might otherwise entail.

Now in a reply to these criticisms of his position (as first presented in Feser 1998a), Lockwood (1998) has conceded that his arguments fail to show that there are qualia of which no one is aware, and support at most the idea that qualia can have *attributes* of which a conscious subject is unaware. But he insists that there is nevertheless reason to believe that there could be evidence in favor of the stronger thesis, and that there is in any case, contrary to what I have suggested, no good reason to suspect the notion of unsensed qualia of being incoherent.

One such piece of evidence, Lockwood suggests, could be provided by a psychological experiment of the following sort. Imagine a subject in a sound-proof room who is asked to listen through headphones to a continuous tone of gradually decreasing volume, to press a button when he can no longer hear it, and finally to report on anything noteworthy that occurs after the button has been pressed. Imagine also that after the button is pressed, the tone continues, unbeknownst to the subject, but that it is eventually switched off, at which point the subject reports that he has detected the change, and even to have detected a phenomenal difference, a difference in qualia, before and after it was switched off. Such a case seems perfectly possible, Lockwood says, and it would provide evidence for his claim that there could be qualia – and not just *attributes* of qualia – which exist unsensed; for there would in this case appear to be an auditory quale which the subject is not aware of at the point at which the button is pressed (when he can no longer hear it), but the continued existence of which is indicated by the subject's reporting a change when the tone is switched off.

That is, the quale continued to exist, even though the subject was no longer conscious of it.

Lockwood himself grants that this interpretation of such an experiment wouldn't be forced on one. For it might also be interpreted as a case of "Orwellian revision," to use some colorful terminology introduced by Dennett (1991, pp. 116-117) to describe a strange sort of psychological phenomenon for which there is some experimental evidence. That is, it *could* be that the tone's being switched off brought about a false memory of there having been a quale post-button-pressing and pre-switching-off – a "rewriting of history" as it were (hence the label "Orwellian").[6] But he insists that his interpretation can't be ruled out either.

Now obviously, as Lockwood would no doubt acknowledge, this sort of experiment cannot have decisive force unless it is actually carried out and has results amenable to Lockwood's favored interpretation (and even then it wouldn't have *decisive* force given the alternative possible interpretations even Lockwood acknowledges). The main problem with the example, though, is that even if it would provide an example of an unsensed quale (rather than merely an unsensed attribute of a quale), it *wouldn't* provide an example of a quale which exists *apart from any conscious subject*, since even if the subject in the experiment wasn't conscious of it,

---

[6] Dennett contrasts this with another psychological phenomenon wherein qualitative features of a perception of a particular kind might be present or absent depending on what other psychological factors are operative, and which he calls "Stalinesque revision" after the Soviet dictator's practice of having photographs and other records "doctored" in order to erase evidence of the existence of some party member who had fallen from his favor; and Lockwood also grants that his three color patch thought experiment, discussed earlier, could be interpreted in *these* terms.

the auditory quale would clearly still be *present* to the subject (as is evidenced by the fact that he notices when it disappears). And it is ultimately *that* sort of quale the possible existence of which Lockwood needs to prove in order to show that the idea that qualia are the intrinsic properties of material objects needn't entail panpsychism.

Moreover, given that, as I've suggested, the burden of proof is on the proponent of unsensed qualia to show that the notion is even coherent, any alternative interpretation of Lockwood's proposed experiment, such as the "Orwellian" interpretation, would seem to be preferable by default. But perhaps part of the point of the experiment (given that Lockwood's imagined result seems at least conceivable) was to support at least the *coherence* of the idea of unsensed qualia, if not their actual existence. And given what I've already said about the experiment showing at most the possibility of qualia which are unnoticed by a subject to which they are (and must be) nevertheless present, perhaps we can grant the coherence of *that* sort of "unsensed quale" (unhelpful as this would be to the realization of Lockwood's main goal).[7]

---

[7] There might be other grounds for this in any case. One interesting piece of supporting evidence (which I thank Galen Strawson for pointing out to me) is provided by Dennett's "Hide the Thimble" example (Dennett, 1991, pp. 334-335), named for a children's game in which a thimble is "hidden in plain sight" and children are asked to find it, many of them being unable to see it even though it is right in front of them, so thoroughly does it "melt" into its surroundings. Even this sort of example, though, might be susceptible of another interpretation. Imagine a case where the thimble is hidden against a background of wallpaper covered with thimble images. Now when an observer fails to spot the thimble in this case, it *might* be that there indeed is a particular "thimble-ish" quale in his visual field that is caused by the thimble (and not by the wallpaper images), and he simply fails to pick it out, thus failing to perceive the thimble. But it may instead be that in this circumstance, there is in fact no particular quale caused by the thimble itself, the wallpaper acting like "static" which prevents the brain from registering, even unconsciously, the actual thimble, and thus prevents it from producing a quale corresponding to the actual thimble (even though other "thimble-ish" qualia are produced). If so, then it's not the case that in failing to see the thimble, the observer has a quale in his visual field of which he is unaware. And the point is that if the "hidden thimble" and similar

Lockwood, though, wants to argue that the coherence of even the stronger notion of "unsensed qualia," on which his position depends, should be uncontroversial. He takes issue with my claim that his acknowledgement of the fact that no one could have the concept of phenomenal red without being able to conceive of what it would be like to be aware of it undermines his claim that qualia can be conceived of as unsensed. For his point was, he says, merely that "in the absence of appropriate sensory stimulation, one can only grasp what phenomenal red is like by *visualizing* it" (1998, p. 417). That is:

> In my view, the reason why only subjects who are capable of sensing or visualizing phenomenal red can grasp the intrinsic nature of phenomenal red is not, as Feser alleges, that the concept of phenomenal red is incapable of being coherently detached from the concept of its being sensed, but that it is only by sensing or visualizing phenomenal red that subjects can become *acquainted* with its intrinsic character. (1998, p. 417)

---

examples can all be accounted for in terms of the latter sort of story (i.e. in terms of background static preventing the production of the relevant qualia), there would be no need to posit "unsensed" quale even in the weak sense under consideration. But I grant that this is a tendentious suggestion, and one that it is not clear is even prima facie plausible in every case. For instance, Searle (1992, pp. 164-165), in a different sort of context, discusses the example of "unconscious pains," e.g. a pain in one's back which, since it can wake someone up even though (being asleep) he isn't conscious of it, would seem to be a quale of which the subject having it isn't conscious; and it's not clear how the "background static" sort of alternative explanation I suggested in the thimble case could be applied here. In any case, again, I needn't be able to rule out the possibility of a quale of which the subject having it is unaware in order to support my main claim against Lockwood, which is that there couldn't be a quale which is not only unnoticed, but exists *altogether apart from* any conscious subject. (And Hayek (1952, pp. 23-25), whose position I'll be defending later on, seems himself to support the possibility of the weaker sort of "unsensed qualia.")

So even though "I can entertain the concept of what phenomenal red is like in itself only... by simulating vision in my imagination," nevertheless, Lockwood claims, "it seems to me that I can readily perform the abstraction that allows me to focus, mentally, exclusively on the phenomenal *object* of this simulated sensing" (1998, p. 417). Just as a naïve realist might suppose that the redness he sees in a physical object exists in the object when he isn't looking at it, even though he can't conceive of it without imagining it visually, we should, Lockwood insists, think of phenomenal red in the same way.

Lockwood's talk of *visualizing* phenomenal red rather than *conceiving of it as being sensed* strikes me as a distinction without a difference. If I first try to visualize phenomenal red, and then try to conceive of its being sensed, it's hard to see how I haven't been doing the same thing both times. Lockwood might respond by saying that the same could be said of visualizing a tree as opposed to conceiving of perceiving a tree, though it would be absurd to suppose that since it seems I'm doing the same thing in both cases, trees cannot exist outside my perceptions – indeed, this is no doubt the point of his accusation that those who reject his view as incoherent are committing the same fallacy Berkeley did. But the analogy between the cases is illusory. For by "visualizing a tree," what is obviously meant is visualizing or conceiving of the appearance of a tree. And as Lockwood himself would acknowledge, in conceiving of the *appearance* of a tree, we're not *really* conceiving of the tree itself in any case, but only of the constellation of qualia it causes in us when we perceive it. To conceive of

the tree itself, so far as we are capable of doing so, would involve instead conceiving of a certain kind of abstract causal structure of the sort physics reveals to us. So the Berkeleian conclusion that trees cannot exist apart from our perceptions is blocked, since we're perfectly capable of conceiving of a tree – as something that has no resemblance, as far as we know, to what we are directly aware of in perceiving a tree – as existing apart from our perception of a tree. But conceiving of phenomenal red really is nothing more than conceiving of its appearance.[8] So there's nothing to block the inference that *it* can't exist apart from our sensing of it, since we can't conceive of it except as it appears to us when we sense it.

Now it's true that we can indeed *talk* about a quale *as if* it were separable from an act of sensing, and even consider it apart from the latter; but this fact does nothing to support Lockwood's view. For we can also talk about, say, a physical object's weight, or height, or even color (either in the objective sense of physics or the naïve realist's sense, for that matter), *as if* they were separable from the object and can consider them apart from the object itself. But they couldn't possibly in fact exist apart from it, and our being able to perform such an act of abstraction doesn't itself suffice to show otherwise. We still can only imagine existing weight, height, or color as the weight , height, or color *of* some particular object; and Lockwood has said nothing to show that phenomenal red, say, even if we can *consider* it apart from the act of sensing it, can *exist* apart from some such act.

---

[8] On Lockwood's own view it is, anyway; though as we'll see in the next chapter, this isn't, after all,

Moreover, Lockwood's appeal to the analogy with naïve realism I think undermines rather than supports his case. For of course, part of the reason why many (including Lockwood himself) reject naïve realism is precisely because we can have no grounds for thinking that external objects have in themselves the properties our perceptual experiences present them as having. Whenever we think of phenomenal red, on the indirect realist view, it is always a feature of our experience that we're *really* imagining, and never a genuine feature of an external object.[9] On the indirect realist view, even though it *seems* otherwise to the naïve realist, he is simply mistaken. So it is odd for Lockwood, who, again, endorses indirect realism, to appeal to the naïve realist view of things to provide a supporting analogy. For if we can't make sense of red physical objects having in themselves, apart from our experiences of them, the phenomenal redness they appear to have, why should we suppose that we can make sense of phenomenal redness itself existing apart from our experiences of it?

Furthermore, even if Lockwood could make out a distinction between being unable to conceive of a quale apart from visualizing it, and being unable to conceive of a quale apart from sensing it, this might still leave him with a problem. For even in that case, he has to admit that qualia can only be conceived of by doing something

---

quite true.

[9] This doesn't contradict the point I made in the last paragraph. I said there that we can't conceive of weight, height, or color existing apart from some object that has them, and this is true even if the qualia associated with these properties aren't properties of physical objects in themselves. For if, by "red," say, we mean the objective physical features that cause red qualia in us, then those objective physical features can't exist apart from some physical object which has them; and if we mean instead the red qualia themselves, then we still can't conceive of them apart from physical objects in the sense that we can only conceive of them as being instantiated along with certain other phenomenal features,

which only a sensing subject can do, which isn't true of things which uncontroversially *can* exist apart from any subject (i.e. I can conceive of a truck, for example, whether or not I've ever had experiences of anything remotely like a truck). And this provides at least indirect support for the claim that we cannot conceive of qualia except as present to – or remembered by – a conscious subject, whether or not the subject is always conscious of them.

Finally, nothing in what Lockwood says has any tendency to undermine my earlier point that qualia seem inherently perspectival, their characteristics determined by a given point of view – a point which seems to support the claim that qualia cannot be conceived of apart from some subject who has them. So Lockwood has failed to undermine the claim that the burden of proof lies on him to show that the notion of unsensed qualia is even coherent; and he hasn't, in any case, met that burden.

I conclude that there *cannot*, after all, be such things as unsensed qualia. It follows that if the RIT is committed to the claim that qualia are what "flesh out" the causal structure of the physical world external to the mind, that is, that they are the intrinsic qualities of the objects making up that world, then it is committed also to a kind of idealism or panpsychism. For if qualia can only exist while present to some conscious subject or mind, then the external "physical" world would then have to be *filled* with such minds! But this, surely, is absurd.

---

such as extension, which are definitive of physical objects as common sense – which takes for granted that such objects really are as they appear – conceives of them.

(b) Panpsychism

Or is it? At least some sympathizers with the Russellian view think otherwise. Chalmers, for one, does, and he in fact (tentatively and with qualifications, to be sure) *embraces* the panpsychist consequences of Russellianism without unsensed qualia as at least a possible, serious solution to the qualia problem (1996, pp. 298-299, 305). On this approach, accepting the Russellian view that qualia are what flesh out the causal structure of the physical world may entail accepting the idea that there are *conscious subjects* associated with these qualia, even in parts of the physical world other than human and animal brains. More precisely, the view is, roughly, that there are what might better be called "proto-qualia" or "proto-phenomenal properties" which flesh out the causal structure of the world and into which qualia of the sort we're familiar with in introspection can be discomposed; and associated with these are "proto-subjects," not conscious minds having anything like the cognitive complexity or experiential richness of ours, to be sure, but conscious minds or "proto-minds" all the same.

This suggestion has opened Chalmers up to some pretty heavy abuse (e.g. Searle 1997) – understandably so, since the position is certainly a radical one. Still, even if understandable, such abuse isn't really fair: Chalmers would defend himself by arguing that taking qualia seriously is bound to lead us to *some* radical reconception of the world or other, a point of view with which I certainly sympathize. Moreover, he might say with some justice that abuse is really all that opponents of panpsychism ever

offer in response; actual *counterarguments* are rare. What I want to do now is precisely to offer some counterarguments, ones intended to show that panpsychism is, not only counterintuitive, but also implausible.

Consider first that on the view under consideration, at least as Chalmers has (tentatively) suggested spelling it out, the character of qualia of the sort we are aware of in introspection is to be explained in terms of more fundamental proto-qualia or proto-phenomenal properties, and ultimately in terms of those proto-qualia which flesh out the most fundamental causal structures described by physics – the proto-qualia which play the roles described by physics' description of fundamental particles (1996, pp. 305-308). And this would entail, presumably, that there are some proto-subjects at the level of fundamental particles which "sense" or have these proto-qualia: Perhaps we are to think of fundamental particles, with their extremely simple functional organizations (and thus very basic information processing capacities) as like little minds (reminiscent of Leibnizian monads), the "conscious experience" of which consists entirely of the having of a single proto-phenomenal property.[10] This isn't necessarily as wild as it sounds. We're not talking about anything that could be

[10] There may be another reason for the Russellian to posit such proto-subjects, apart from the failure of Lockwood's attempt to show that there could be unsensed qualia, and that is that if we don't regard qualia or proto-qualia as associated with proto-subjects, we would presumably have to treat them as *particulars* in their own right. And as noted earlier, though qualia are sometimes spoken of by philosophers as if they *were* particulars, they are more commonly and explicitly treated as properties; and for good reason, since they surely *seem* to be properties, namely properties of experience (and, of course, of external objects, though only erroneously so, on the view defended in this essay). Moreover, there are special problems associated with trying to conceive of *proto*-qualia (as opposed to qualia) as particulars: We'd have to think of them as in some way "particle-like," and as taking up a definite position in space; and even if this seems at least vaguely plausible if we think of them on the

described as remotely like a *person* or as having conscious experiences at all remotely like ours. If we avoid the (admittedly almost irresistible) urge to think of all this in anthropomorphic terms, it can be argued that the suggestion that such particles have something *remotely like* experience need be no more radical than the suggestion that they have something *remotely like* shape, etc. Nevertheless, this suggestion is highly problematic.

Take two brain states, A and B, and assume, with the RIT, that they're associated with two different qualia, a reddish color quale and a pungent odor quale, respectively. On the suggestion under consideration, the qualitative character of each is to be accounted for in terms of lower level proto-qualia. Now John Foster (1991, pp. 127-128) argues that on the Russellian view, the fundamental particles composing a brain event identical to a pain, say, would have to be composed of something like "pain particles," in which case, since all fundamental particles would have to be of the same sort, the Russellian would have the problem of explaining how a brain state with a different phenomenal character – a color quale, for instance – could also be composed out of "pain particles."[11] But though Foster is, I think, right to say that fundamental particles must be of the same general character, it isn't so clear that

---

model of color patches, say, it is very odd and implausible if we think instead in terms of tastes, odors, or sounds.

[11] Foster argues thus in the course of developing, against the Russellian view, an objection based on what is generally referred to as "the grain problem," which is the problem of explaining how it can be that qualia seem so much "smoother" and less finely grained in structure than brain states do if they are *identical* to brain states (and is thus a problem for any version of the identity theory, not just the RIT). (Lockwood (1989, pp. 16 and 177, and 1992) takes this to be the most serious problem for the

accounting for the character of a quale in terms of proto-qualia has to go the way he thinks it does. After all, there's no need to suppose that objects as different in shape as chairs and tables need be accounted for in terms of fundamental particles of different sorts: We needn't suppose that chairs are composed of "chair particles," so that there's a problem of explaining how tables could also be composed of chair particles. Just as higher level properties of different sorts of physical objects can be accounted for in terms of the various *combinations* that particles of the same sort can take, so too, perhaps qualia can be accounted for in something like a combinatorial way. The Russellian can thus assume that proto-qualia are all of the same (proto-) phenomenal character (a character somehow neutral as between the characters of the various qualia we're familiar with from introspection).[12] And he can go on to say that a certain combination of proto-qualia will yield a quale like that associated with A, and another combination will yield a quale of the B sort.

Nevertheless, it is difficult to see how even combinatorial differences could account for the differences between the qualia associated with A and B. For it is not only possible, but highly likely, that brain states highly alike in structure, considered by themselves, are associated with different qualia. For any micro-level at which it might be claimed that there is a proto-qualia combinatorial difference – which would surely

---

RIT, though I think the problems I've been discussing are at least equally serious, if not more so.) I'll say a little more about this problem in the next chapter.

[12] Indeed, he will have to assume this anyway, given a thesis of Chalmers' that we'll have reason to look at later on, namely the *principle of organizational invariance*, which entails that all protons, say, if alike in functional organization, would have to be associated with the same proto-phenomenal content.

require a difference in physical structure – adequate to account for a difference in associated qualia, it seems possible and likely that there are brain states associated with different qualia which do *not* have the (allegedly necessary) differences in structure. If we start at the bottom level of fundamental particles and work upward, it seems we'll never get to a combinatorial level where it is plausible that all and only those brain states having the physical structure subserving the given combination of proto-qualia will have just the quale in question. And if we start at the top level and work downward, it seems like we'll end up having (implausibly) to associate different (proto-)phenomenal contents with combinatorial structures in A and B which are similar, if not with the (structurally identical) fundamental particles making up A and B.[13]

In short, trying to explain the phenomenal content of qualia in terms of proto-qualia appears to be a non-starter. So panpsychism doesn't have anything like the explanatory advantages it might be thought to have, and which, it is claimed, justify us in accepting it despite its eccentricity.

Even aside from its explanatory deficiencies, however, the very idea of proto-phenomenal properties associated with proto-subjects is highly dubious. Consider that in order even to get a grip on the concept of such proto-phenomenal properties, we'd surely have to think of them on the model of the qualia we're familiar with in everyday experience. But it is difficult to see how we can do so given that, as noted above,

[13] Not only is this prima facie implausible, but it would, again, violate the – extremely plausible, as

we'd have to conceive of something neutral as between the various qualia, abstracting away anything like reddishness, pungency, loudness, coolness, bitterness, etc. etc. – in which case it's hard to see what residue would remain to serve as a common underlying basis for all phenomenal character. Moreover, one defining feature of qualia seems to be what I have called their *perspectival* character: The character had by the reddish quale I have when I look at a tomato is determined partly by the fact that I'm looking at the tomato from a particular *point of view*; and in general, it is difficult to see how we can conceive of qualia apart from the point of view of some subject. But what sort of "point of view" are we to associate with the proto-subjects with which proto-qualia are to be associated? There's nothing remotely like the factors that determine a unique point of view in the case of human beings and other animals – sensory organs, a unique position among significantly varying physical objects, etc. – in the case of a proton, say: A proton has nothing like eyes which provide, but also delimit, a field of vision; it is not surrounded by objects which significantly differ from itself and from one another and which would thus make for a contrast between the "sights" which would greet it when it turned this way or that; etc. So it seems difficult, if not impossible, even to get so far as to conceive of what it is we're talking about when we talk about proto-qualia.

At any rate, given the difficulties involved in panpsychism, the burden of proof is on its advocate to show both that it can be made sense of and that it provides us

---

we'll see – principle of organizational invariance.

with any explanatory advantage. That burden has not, in my view, been met. And that the RIT seems inevitably to entail panpsychism thus constitutes strong evidence against it.

(c) Russellian zombies?

There's worse to come. I noted earlier the respects in which the RIT is alleged to be immune to the objections typically made to the mind-brain identity theory. Closer examination, however, reveals in my view that it is, after all, no less subject to some of those objections. Consider first how the zombie argument might be used against the Russellian.

The Russellian response to the zombie argument was, we saw, to argue that to imagine a world of creatures having just the neurophysiology we have would just be to imagine a world having a certain kind of causal structure, and would not by itself involve imagining what it was that fleshed out that causal structure. The upshot was that the possibility of imagining such a world devoid of qualia – a zombie world – was not a threat to the RIT, since that theory never claimed that to imagine a world of creatures with brains *exactly like ours* it is sufficient to imagine all and only the neurophysiological facts being exactly the same. One would *also* have to imagine qualia as what fleshes out the neurophysiological structure, since they're what do so in the actual world, according to the RIT; and in doing so, one would, of course, not be imagining a zombie world at all.

But this just raises a further question. If what I know in knowing all the neurophysiological facts is just the abstract, causal structure of the brain, and not what intrinsic features flesh out that causal structure, then what reason is there to suppose that what fleshes it out *must* in fact be qualia? If all that there is to being a certain kind of brain event is playing a certain causal role, why couldn't any number of things, including things lacking the features definitive of qualia, play that role? There seems no reason at all to think that nothing else could. But if something else could, then it follows that it is possible for all the facts about the brain *qua brain* – that is, all the facts which make something a brain, which would be neurophysiological facts about a brain's causal structure – to be just as they are, and yet there are no qualia. And if this is possible, then a zombie world is possible. But if the mind just is the brain, then a zombie world shouldn't be possible. It follows that the RIT, which identifies the mind with the brain, is false. (The same point might also be put – and perhaps be put even more clearly – in Kripkean terms: the claim that the mind is identical to the brain, even if understood in terms of the RIT, cannot, if true, be false in any possible world; but there clearly are possible worlds, i.e. those where something other than qualia fleshes out the causal structure of the brain, where it is false, so the claim is false simpliciter.)[14]

[14] Similar reformulations could obviously be made of some of the other objections we've looked at as well, e.g. the knowledge argument: Mary could know all there is to know about what it is that makes the brain the sort of thing it is, i.e. its causal structure, and yet know nothing about qualia.

Now one might be tempted to respond that it is just implausible to suppose that something could play *exactly* the neurophysiological role the RIT takes pain, say, to play – that is, it results from bodily injury, brings about pain behavior, and so forth – and yet lack the qualitative feel of pain, that is, not be identical to a pain quale. But such a response would not sit well with the conceptual dissociation of qualia and functional role that the objections to physicalism we looked at earlier appealed to – objections championed by Russellians themselves.[15]

Moreover, we must keep in mind that, as was pointed out in the last section, just as the structure of higher-level neural processes must be explained in terms of lower-level processes, until we reach the level of fundamental particles, so too, on the Russellian view, must the qualitative character of higher-level qualia be accounted for in terms of that of the lower-level "proto-qualia" associated with fundamental particles. So it is ultimately the plausibility of something other than (proto-)qualia playing the roles associated with protons and electrons, say, on which the objection we're considering stands or falls; and this seems very plausible indeed. As Chalmers himself points out (1996, pp. 135-136), it would be not only implausible, but perhaps even incoherent, to suggest that anything that lacked the proto-phenomenal character of what the RIT might claim in fact plays the role definitive of electrons just wouldn't *really* be an electron even if it played the role in question; fundamental particles seem

[15] Such a response is not without plausibility, however, because the functionalist linking of functional role and qualitative character is itself not without plausibility, despite the failure of physicalist varieties of functionalism. In the next chapter, I will try to show that the only genuine solution to the

to be defined *entirely* in terms of their causal roles. But if so then there can be no way to rule out the possibility that something other than proto-qualia do in fact play these roles, and thus no way to rule out the possibility that something other than qualia could play the higher-level roles associated with brain states.[16]

(d) "Neural chauvinism"

There is at least one other important respect in which the RIT is subject to the same criticisms made of physicalistic mind-brain identity theories. The latter sort theory was accused by the early functionalists of "neural chauvinism." That is, it was objected that it was perfectly conceivable that creatures whose physical constitution differed radically from ours (creatures composed of silicon, say) could have exactly the sorts of mental states we have, and that this shows that an adequate characterization of mental phenomena must abstract away from the various ways mental states might be realized. What is important is not the sort of "stuff" in which mental states are instantiated – even neural tissue – but the way that stuff is organized.

I think the same objection applies to the RIT. Why suppose qualia should turn out to be *brain* states in particular, even given that the brain is understood merely in terms of its causal structure? Or, to put the question in a way which is perhaps more appropriate to the RIT: Why suppose that *brain* states alone should turn out to be qualia? Why couldn't, say, certain *silicon* states (understood in terms of abstract

---

qualia problem is precisely one which marries the insights of the Russellian view to those of functionalism.

causal structure, of course) turn out to be identical with qualia as well? Even on the RIT, the same functionalist point seems to apply, namely that an adequate characterization of mentality must abstract away from neural tissue, silicon, etc. – even if these are conceived merely in terms of causal structure – so that the multiple realizability of mental phenomena can be accounted for. And it is hard to see how the RIT can provide such a characterization. It's only resource for doing so would be to make qualia out to be what fleshes out the causal structure of *every* physical system, which, as we've seen, entails a commitment to panpsychism – in which case it wouldn't be "chauvinistic" *enough*, ascribing mentality to systems even the functionalist would deny have it.

This consideration leads us, quite naturally, to the point where we are ready at last to develop what I take to be the most satisfactory solution to the qualia problem, a solution which incorporates the enduring insights of the Russellian view which, as I think I've now shown, cannot finally be judged to be successful. For on this solution, it is precisely its preservation of the "neurocentric" aspect of the materialist legacy which keeps the RIT from being a fully adequate account of the place of mind in the natural world. It is to my mind precisely the predominantly *functionalist* tenor of contemporary philosophy of mind which constitutes its lasting contribution, just as it is it is its recognition of the irreducibility of the first-person subjective realm (as

---

[16] This is no doubt one reason why Chalmers' commitment to the Russellian view is more tentative than that of some of its other proponents – and also (as he hints at, 1996, p. 136) probably why, as noted earlier, he classifies his view as a kind of property dualism rather than an identity theory.

embodied in its commitment to indirect realism) – while at the same time rejecting dualism – which constitutes the lasting contribution of Russellianism. The combination of these of these elements is, I will argue in the next chapter, the key to as complete a solution to the qualia problem as we are likely ever to have. And yet, such a combination has, to my knowledge, never even been considered in the history of the philosophy of mind – save perhaps in a few suggestive but obscure hints in some neglected writings of the economist, political philosopher, and sometime-philosophical-psychologist F.A. Hayek.

6. Hayek and functionalism

(a) The project of The Sensory Order

The Vienna of the late 19th and early 20th centuries must rank with ancient Greece and Enlightenment Europe as one of the great fountainheads of Western cultural achievement. The list of great names associated with that time and place is evidence enough for this claim: Sigmund Freud, Gustav Klimt, Arnold Schönberg, Ernst Mach, Moritz Schlick, Rudolf Carnap and the rest of the Vienna Circle, Ludwig Wittgenstein, Karl Popper, Ludwig von Mises (brother of the logical positivist Richard von Mises) and other members of the Austrian School in economics, are just a few. Friedrich August von Hayek is another of these great names, and Hayek's roots in this milieu run deep. He was a student of Mises and was himself the most famous and influential member of the Austrian School, a cousin of Wittgenstein, and a close friend and collaborator of Popper's; and as a student he participated fully in the interdisciplinary intellectual life of the time.[1] Deeply influenced by the work of Mach and Schlick, he had originally intended a career in theoretical psychology, and had in 1920 written up some sketchy, original ideas of his own on the subject. But events took him instead down the path of economics and, ultimately, political philosophy (and, geographically, to England and later the United States), his work in these fields winning him fame and, in the case of economics, a Nobel prize in 1974.

---

[1] See Hayek (1994), Smith (1994), and Gray (1998) for more detailed discussions of Hayek's relationship to this Viennese context, as well as, more generally, to other parts of the 20th century intellectual context.

The dramatic failure of Keynesian economic policies had by that time made Hayek's defense of the free market once again popular, but things were very different at the time Hayek published his best-known book, 1944's semi-popular The Road to Serfdom, which did not endear him to the intellectual establishment of the day. Then Keynes had been riding high (though Hayek was up until then an influential rival), and the book's attack on the socialism that still held so many Western intellectuals in its thrall was nearly fatal to Hayek's career.[2] Unable for a time to find a position in his own field, he took up a post at the University of Chicago's Committee on Social Thought. Stung by the reception of his peers to The Road to Serfdom and determined to take a rest from controversies in economics and public policy, he dusted off his 1920 manuscript and turned again to the problems of theoretical psychology. The result was 1952's The Sensory Order.

---

[2] Rudolf Carnap, for example, chastised Popper for praising Hayek's book – though he admitted that he himself had not read it (Hayek 1994, p. 17)! As great a figure as Carnap was, such a lapse in intellectual integrity is not completely surprising given that, according to his student and friend Hilary Putnam, Carnap "felt strongly that for all x, planned x is better than unplanned x. I know this from conversation with him, and it is also evident in his intellectual autobiography. Thus the idea of a socialist world in which everyone spoke Esperanto (except scientists, who, for their technical work, would employ notations from symbolic logic) was one which would have delighted him" (1994, p. 185). Like socialism, the logical positivism which Carnap championed was a variety of the "constructivist rationalism" (Hayek, 1997) which Hayek attacked throughout his career, which he took to be naïve and even dangerous in its influence on moral and political thinking, and which he contrasted with the kind of "critical" or "evolutionary" rationalism he shared with Popper and which we'll have reason to look at later. Interestingly enough, however, Schlick, Carnap's fellow logical positivist, would likely have taken Hayek's side had he only lived long enough. As Herbert Feigl reports, though "Schlick in his early years had been sympathetic to the ideals of a pacifist socialism... the rise of Nazism in Germany, among other factors, impelled him to modify his outlook in a more conservative and individualistic direction... [M]y impression... was that he was deeply shaken by the events in Germany and that he no longer maintained as steadfastly as before his belief in 'salvation' through human kindness" (page xv of the introduction to Schlick, 1985). Schlick's disillusionment with the notion of human perfectibility was, of course, tragically prophetic: he was later murdered by a deranged student with National Socialist sympathies (Monk 1990, p. 357).

That book, as I've said, contains hints at a solution to the qualia problem which combines insights from the Russellian view and from functionalism, but which avoids the apparently insurmountable difficulties facing those views. That those hints have gone unnoticed is no doubt primarily because the book, written by someone outside philosophy, strictly speaking (not to mention outside psychology and neuroscience) has simply not been widely read – though it has received some attention from Hayek scholars mainly interested in the light it might throw on Hayek's economics and social philosophy (Miller, 1979; Fleetwood, 1995; Gray, 1998), and even from a few cognitive scientists and neuroscientists who have seen in it prescient insights foreshadowing the connectionist or neural network paradigm so widely discussed in current artificial intelligence research (Fuster, 1995; see also Smith, 1996), as well as such related approaches as "neural Darwinism" (Edelman, 1982, 1987) and the "complex adaptive systems" research associated with the Santa Fe Institute (Miller, 1996).[3] But another likely reason is that Hayek does not frame the issues in the way they tend to be framed today, a way which in my view has made the qualia problem clearer than it previously has been in the history of philosophy; and his project was thus not precisely that of "solving the qualia problem." So he never addresses the issues of whether facts about qualia supervene on physical facts, whether

[3] There has been some philosophical attention paid to it as well, though not much. Hamlyn (1954) and Sprott (1954) reviewed the book for two of the most important journals in Anglo-American philosophy, but even so, the only other studies of the book from a philosophical viewpoint that I know of are Agonito (1975), Weimer (1982), de Vries (1994), Dempsey (1996), and the already mentioned Gray (1998) and Smith (1996). And none of these addresses, as I aim to, the way in which Hayek's views might open the way to a solution to the qualia problem.

zombies are logically possible, and so forth – issues of the sort we are primarily concerned with in this study.

We would do well, then, to get clear on exactly what Hayek *was* up to in The Sensory Order. Only after spelling out the problem that book was concerned with and the solution Hayek proposed will we be able to see how his work might apply to the problem with which we are concerned.

Under Mach's influence, Hayek became convinced that all we are ever directly aware of in perception are sensory qualities or qualia.[4] Unlike Mach (who advocated a kind of neutral monism), however, he did not think that qualia were all that existed. Like advocates of the Russellian view, he was a scientific realist of sorts in that he acknowledged the reality of a physical world external to the mind, a world which (as Russell also famously argued) ultimately consists not of *objects*, but *events*.[5] But, again in line with Russellianism, he believed that we know this world only indirectly, via our direct awareness of qualia, and that what we do know of it is only its *structure*.[6]

---

[4] As with so many of the other writers we've been looking at, Hayek did not himself use the term "qualia," but almost always spoke of "sensory qualities." Nevertheless, for uniformity's sake, I will continue to use "qualia" when discussing his views.

[5] This is true, in Hayek's view (and Russell's, for that matter) also of the internal world of the mind; for him the mind is not a kind of substance but rather a system of events. Thus he tends to speak of mental *events* rather than, as I've generally done here, mental *states*. But the substance (if you'll pardon the pun) of his views is more or less the same whether we describe them in terms of states or events. (Likewise, despite Russell's own stress on an event ontology, Russellians like Lockwood tend to talk of mental states rather than mental events, if only because such talk has become fairly standard in philosophy of mind. There's no problem with this so long as there's nothing said which implicitly commits one to a *substance* ontology of some sort.)

[6] Indeed, it is very likely that these similarities with the Russellian view are not coincidental. For not only was Mach also an influence on Schlick (a pre-Russell "Russellian") and Russell themselves, but

It was the precise relationship between the realm of qualia, or "the sensory order" of the mind, as he called it, and the physical order of the external world as revealed by physical science, that was Hayek's concern. For science has revealed to us that the latter world is very different from the way it is presented to us by our senses in the former world. In particular, events that appear to be of the same sort in the sensory order are revealed by physics to be very different, and events that appear to be very different as they are presented to us by the senses are shown by physics to be similar: for example, "the sensation of 'white'... can be produced by an infinite variety of different mixtures of light rays" (1952, p. 14), while "the same vibration which, if perceived through the ear, will be experienced as a sound, may be experienced as a vibration by the sense of touch" (1952, p. 13). So how exactly does the one world give rise to the other? How do events in the objective world described by physics get represented in the subjective realm of the mind in exactly the way they do? (To use the terminology made famous by Wilfrid Sellars: How do events in the world as described according to the "scientific image" bring about the events familiar to us in the "manifest image"?)

What we know of the external, objective physical world is, again, for Hayek no less than for Russellians, just its structure, that is, the relations that hold between events taking place within it as described in the mathematical language of physics. But

---

Schlick, as already noted, was a direct influence on Hayek, and Russell's work was also read by Hayek, as the bibliography of The Sensory Order indicates – though Hayek seems to have counted Schlick as the greater influence and never makes special mention of Russell as having influenced his views.

on Hayek's view, something similar is true as well of our knowledge of the internal, subjective mental world. What we know of the sensory order are just the relations existing between the events taking place within it. Now the principles governing the relations between mental events famously suggested by Hume were resemblance, contiguity, and cause and effect. Hayek's own emphasis is on the first of these, and in particular, he stresses the features shared by qualia associated with different sensory modalities, such "intermodal" attributes as those of being cool, warm, strong, weak, mild, mellow, tingling, and sharp, in regard to which, Hayek says, "we are often not immediately aware to which sensory modality they originally belong" (1952, p. 21).[7] It is precisely in terms of such relations that we typically describe qualia: if asked to describe the blue quale I'm having right now, I have to do so by saying such things as that it's similar to the qualia I have when I look at the sky and the sea, somewhat less similar to the qualia I have when I look at grass, but more similar to those than to the qualia I have when I look at a "Stop" sign, similar in one respect – "coolness" – to the qualia I have when I pick up an ice cube and not to those I have when I put my hand over a flame, etc. etc. Beyond the description I can give them in terms of their relations to other qualia, as well as relations to sensory inputs, behavioral outputs, and other kinds of mental states, it seems that I can say nothing at all about them: "[A]ll

[7] The surprising similarities existing between various sensory modalities is further evidenced by cases of "prosthetic vision," wherein blind patients are enabled to have experiences very much like visual experiences through a device which is attached to the back and stimulates the skin in response to signals sent from a television camera (Hofstadter 1981, p. 411). This sort of example indicates that the distinction between touch and sight, and between the qualia associated with each, is not as rigid

that can be communicated are the differences between sensory qualities, and only what can be communicated can be discussed" (1952, p. 31). Indeed, this is precisely the reason qualia are often said to be "ineffable." But of course, describing them as ineffable makes it sound as if there were more to be said about them – because we know more about them – than just their relations; and Hayek denies this:

> It seems thus impossible that any question about the nature or character of particular sensory qualities should ever arise which is not a question about the differences from (or the relations to) other sensory qualities; and the extent to which the effects of its occurrence differ from the effects of the occurrence of any other qualities determines the whole of its character. (1952, p. 35)

As this passage indicates, on Hayek's view, the relations of similarity or dissimilarity between qualia themselves ultimately turn out to be (or are analyzable in terms of) causal relations; he seems to mean, specifically, that a given quale's similarity or dissimilarity to another quale amounts to a tendency to evoke or a tendency not to evoke, respectively, an instance of that other quale (or, say, a memory of it). Moreover, the whole of a quale's character is determined by the *entirety* of the relations that exist between events in the sensory order: "[A]ll mental qualities are so related to each other that any attempt to give an exhaustive description of any one of them would make it necessary to describe the relations existing between all" (1952, p. 23).

---

as appears at first glance; and Hayek is suggesting that the same is true of qualia in general, in that

In short, Hayek's view, despite its similarities with the views of Mach and the Russellians, differs from them in denying that there are any "absolute" features of qualia – that is, in denying that qualia are *intrinsic* properties.[8] Our knowledge of the internal world, as well as of the external world, is knowledge not of intrinsic nature but only of structure:

> [I]f we can explain how all the different sensory qualities differ from each other in the effects which they will produce whenever they occur, we have explained all there is to explain... [T]he whole order of sensory qualities can be exhaustively described in terms of (or 'consists of nothing but') all the relationships existing between them. There is no problem of sensory qualities beyond the problem of how the different qualities differ from each other – and these differences can only consist of differences in the effects which they exercise in evoking other qualities, or in determining behavior... The conclusion to which we have been led means that the order of sensory qualities no less than the order of physical events is a relational order – even though to us, whose mind is the totality of relations constituting that order, it may not appear as such. The difference between the physical order of events [external

---

they are all ultimately describable in terms of one another.

[8] As we'll see, this is not to say that there is *nothing* here which has any intrinsic properties. Hayek, like Russell, would identify qualia with features of the brain, and the brain itself no doubt has intrinsic properties of some sort or other. The point is rather that these intrinsic properties are not what is known when one introspects qualia. The brain state one introspects when one introspects a quale has intrinsic properties, but its qualitative character is not among those properties; rather, the brain state has that character only by virtue of its relations to other brain states. Put another way, the quale, *as a quale*, has no intrinsic character, its nature as a quale being determined entirely by its

to the brain] and the phenomenal order in which we perceive the same events is thus not that only the former is purely relational, but that the relations existing between corresponding events and groups of events in the two orders will be different. (1952, pp. 18-19)

The way this fits in with Hayek's project of explaining the way the objective, physical order gives rise to the subjective, sensory order is this. Part of what we know about the relationship between the two orders is that latter depends in some way on what takes place in a part of the former, namely in that portion of it studied by neuroscience. The brain, that is, is a part of the larger physical world, and the mind is in some way related to it; indeed, for reasons we've already looked at, it seems that there is good reason to *identify* the mind with the brain. We also know that events in the brain are brought about by events in the rest of the physical world in the manner revealed by the physiological study of perception. So since we can, in principle, understand the way in which the rest of the physical world brings about the events taking place in the brain, we can also in principle understand the way in which it brings about the representation of itself that constitutes the sensory order. We can do so if we can discover a way in which the structure of relations between events constituting the sensory order and the structure of relations between events constituting the neural order (at some relevant level of description) are *isomorphic*. For if we could know, through neuroscience, how the physical world produces the latter sort of system of

---

relations – even though, considered merely as a brain state, as the thing which has the relations in

relations, then we would know how it could produce the former sort of system, since the two would be identical in structure, and *all we know* of either is structure.

The bulk of The Sensory Order consists of spelling out in detail a neurophysiological account of perception on which the structure of the neural events constituting the brain can indeed be seen to be isomorphic with that of the mental events making up the sensory order. In particular, Hayek focuses on the way in which neural events, by virtue of the relations existing between them, serve to *classify* events in the external environment in a way that parallels the classification that the senses perform.[9] As the external world impinges on an organism's sensory surfaces, connections are formed between neurons and groups of neurons in such a way that the impulses they carry occur together in a regular way. Different groups of impulses come to be associated with different features of the external world as those features produce different sets of neural connections in the brain. It is not the intrinsic character of any impulse or group of impulses that gives rise to a correlation with a feature of the external world; rather, it is "the position of the individual impulse or group of impulses in the whole system of such connexions which gives it its distinctive quality" (1952, p. 53) – just as it is the position of a given quale within the system of

---

question, it does have an intrinsic nature (which nature is unknown to us).

[9] "Classify," of course, has intentional overtones, and thus it may be thought that the use of this concept is question-begging if intentionality is part of what Hayek's account is meant to explain – as indeed, to an extent, it is, as we'll see in the next chapter. But Hayek is not using the term in an intentional sense. He has in mind a technical sense in which it pertains merely to a system's differential reactions to events taking place outside it, in particular its going into various kinds of states. "Classify," given its cognates "classes" and "classification," is most useful in describing the

mental events which gives it *its* distinctive quality. In this way, the brain "classifies" features of the external (and internal) environments; moreover, higher-order connections come to be formed which effect a multiple classification of features, that is, the features come to be associated with different groups of neural impulses and thus fall into more than one "class."

In this manner, the external physical world brings about an order in the brain which mirrors in its structure the sensory order of the mind (and is in fact identical to that order). Hayek is perhaps even clearer in spelling this out in "The Primacy of the Abstract" (1978), a later essay. In perceptual experience, he there says, impulses within different sets of neural connections are initiated by different aspects of a given stimulus, some sets of connections associated with some properties, others with others. That I see something as an object of a certain sort, and respond behaviorally to it in a certain way, is the result of a "superimposition" of the members of one set of neural events and dispositions to act rather than another (1978, pp. 40-42). The superimposition being of the sort it is is what gives the set of neural events and dispositions to act constituting it the sensory character it has. Using a simple example, we can illustrate what Hayek has in mind. Consider the case of my looking at an orange. What gives this experience the quality it has, a quality which is similar in some respects but not others to that of the experience of looking at an orange car, is that the orange's stimulating my sensory organs initiates some sets of neural impulses which

---

sort of phenomenon he's interested in, though he grants that, to avoid misunderstanding, "grouping"

are also initiated when I look at an orange car and others which are not, but which are also initiated when I look, say, at a billiard ball (which is similar to an orange in shape); that those impulses initiate further sets of impulses that are related to those initiated when, say, I see other types of fruit (while failing to initiate impulses related, say, to my seeing rocks); and that it ultimately (through such intermediate impulses) initiates some dispositions to act (realized in further neurophysiological activity), rather than others, say a disposition to salivate and to eat the object (which I also have when seeing a hamburger), rather than a disposition to take a drive, which I might have when seeing an orange car. In short, that it is just this collection of interconnected neural impulses rather than another is what makes it identical to a "round-ish, orange-ish" quale rather than, say, a "reddish, square-like" quale.

That a certain set of neural impulses is correlated with a certain property, and that only the superimposition of such a set upon others, correlated with other properties, makes possible the distinctive character that a given quale has, entails, Hayek also suggests, that sensory experience is possible only once one has, in virtue of the development of such connections in his brain, formed concepts of the properties in question (1978, pp. 42-3) – a person's having a concept being identified with his having formed a certain set of neural connections.[10] From this, Hayek argues, it follows that the having of general concepts is a presupposition of experience rather

---

might be a better term (1952, p. 48).

[10] In The Sensory Order, Hayek refers to the pre-experiential development of such connections, confusingly, as "pre-sensory experience" (1952, pp. 165-172).

than being the product of abstraction from what is presented in experience, as classical empiricism would have it (1978, pp. 42-43; 1952, pp. 165-172). This is what Hayek means by "the primacy of the abstract," and it gives his position an obvious affinity with Kant's – though it's hardly the only Kantian element given Hayek's view, which he shares with Russellianism, that we cannot know the intrinsic nature of the external world.

Moreover, the neural connections in question are not just the product of the development of such connections, and thus concepts, that occurs as a result of an individual organism interacting with a particular environment. They are also partly the product of the evolutionary history of the species to which the organism belongs (1978, p. 42; 1952, p. 166). The individual organism is predisposed to form concepts, or sets of neural connections, that have proved advantageous to the preservation of the species; and presumably predisposed not to form those which might somehow prove disadvantageous (as Hayek implies on p. 42 of 1978). The character of sensory experience, and of qualia, is thus partially determined by natural selection. Altogether, Hayek says:

> Sense experience therefore presupposes the existence of a sort of accumulated 'knowledge', of an acquired order of the sensory impulses based on their past co-occurrence; and this knowledge, although based on (pre-sensory)

> experience, can never be contradicted by sense experiences and will determine the forms of such experiences which are possible (1952, p. 167).[11]

The environmental and evolutionary factors Hayek calls attention to bring to mind further affinities not just, again, with Kant – for whom the a priori element in perceptual knowledge was determined by the very possibility of having any experience at all, and thus had a universal character – but also with such 20th century advocates of the "theory-laden" character of perception as Sellars (1956) and W.V. Quine (1950), for whom there is nothing *necessary* and unalterable about the categories which determine the nature of experience. Hayek was also a critic of "the myth of the given," to use Sellars' memorable phrase. His account of perception led him to deny that qualia – or sense-data, a more appropriate description for qualia in their guise as an allegedly indubitable foundation of knowledge – provided some *neutral* epistemological ground by which to adjudicate between theories. Nevertheless, given that his inspiration was a biological one, in that for Hayek, it is neurophysiological and evolutionary factors, rather than merely cultural ones, which ultimately determine the nature of perception – but also *constrain* the possibilities for its alteration – his account does not threaten the sort of relativism that Quine's and Sellars' positions have been thought to entail (however contrary to their own intentions) and which has been argued for explicitly by writers like T.S. Kuhn (1970).

---

[11] Weimer, in an apt phrase, thus characterizes Hayek's view as a kind of "physiological apriorism" (1982, p. 270)

In any case, all this tends further to solidify the notion, already supported by the indirect realism and structuralism Hayek shares with Russellianism, that we lack any knowledge of the intrinsic nature of the external material world. As Hayek puts the point earlier made (somewhat more colorfully) by Russell: "Whenever we study qualitative differences between experiences we are studying mental and not [external] physical events, and much that we believe to know about the external world is, in fact, knowledge about ourselves" (1952, pp. 6-7).

As obvious as the similarities with Russellianism are the differences, though, in particular Hayek's commitment to a kind of *functionalism*. For like the functionalism so popular with contemporary physicalists, Hayek's view is also committed to a conception of the mind as a kind of abstract structure, where mental states are defined in terms of their functional roles rather than in terms of the nature of the stuff in which they are instantiated:

> It is at least conceivable that the particular kind of order which we call mind might be built up from any one of several kind of different elements – electrical, chemical, or what not; all that is required is that by the simple relationship of being able to evoke each other in a certain order they correspond to the structure we call mind. (1952, p. 47)

In particular, the *intrinsic* nature of the elements is irrelevant to their status as mental; anything that played the appropriate role would thereby be a mental item. And, most relevant for our purposes, this would include *qualia*: they too are defined in terms of

relations rather than intrinsic character. Knowing qualia, on Hayek's view, is *not*, after all, knowing the intrinsic properties of the brain. Our introspection of qualia is indeed, as for the RIT, the direct introspection of brain states, but only of those states *qua* playing a certain role; indeed, it might more accurately be said that introspection of qualia is the direct introspection of *functional* states. In knowing a quale I know a given brain state, but *only as* a state which bears certain relations to other brain states; I do *not*, pace the RIT, thereby know the intrinsic character of that brain state.

Functionalism is the view which has given rise to what is often called the *computer model of the mind*, on which the mind is not just an abstract structure, but an abstract structure of the sort represented by a *program*. The mind is best thought of as a kind of "software" which is in the human case run on the "hardware" of the brain, but could in principle be run or instantiated on any number of other kinds of physical system. Now in contemporary cognitive science and artificial intelligence research, there are two competing paradigms of the *sort* of program the human mind represents. The classical *symbolic processing* paradigm models the structure of the mind on the structure of language, mental states being conceived of as formal symbols corresponding to such semantic units as sentences, the relations between the symbols being *causal* relations which mirror the *logical* relations between propositions. The *connectionist* or *neural network* paradigm, on the other hand, models the structure of the mind on the structure of the brain, taking its fundamental units to be sub-symbolic and related to one another, not by causal relations mirroring the semantically sensitive

logical relations between propositions, but rather by causal relations which mirror the (decidedly sub-semantic) excitation and inhibition relations holding between neurons in the brain.

Hayek's functionalism is clearly of the *connectionist* variety; and as a number of researchers have come to recognize, he was in fact one of the (largely neglected) fathers of connectionism, along with the (more widely recognized) D.O. Hebb (1949).[12] As we'll see later, that his is a connectionist rather than symbolic processing approach has important philosophical consequences.[13]

So much, then, for spelling out the project of The Sensory Order. We now turn to the question of how exactly it contains the elements for a solution to the qualia problem, as I have claimed it does. And it might at first not be at all evident how this is so. For despite its commitment to part of the Russellian position, doesn't the functionalist element undermine what I've implied is the most valuable part of that position, in virtue of which it constitutes at least an advance over physicalism – namely its respect for qualia? Doesn't Hayek's account of perception as classification, and thus of qualia as classificatory or discriminative states put him in the same camp as

---

[12] Hayek wrote (1952, p. viii) that he even considered not going through with publishing The Sensory Order when Hebb's book came out as he was polishing the draft, but decided the book would not be redundant given that Hebb's book focused on neurophysiological detail rather than the working out of theoretical underpinnings, which latter task was the main concern of his own book.

[13] It is not clear that *every* aspect of Hayek's position, or even the respect in which, I will argue, it solves the qualia problem, depends on its being a connectionist version of functionalism; and I believe at least the basic idea of qualia being both irreducibly subjective and non-intrinsic or relational can be spelled out on more or less any form of functionalism. In any case, as is well known, one needn't commit oneself exclusively to either the connectionist or the symbolic processing model, and there are models which combine elements from both, so that we needn't decisively settle the controversy

Dennett – whose position we've already found wanting? Doesn't Hayek, no less than Dennett and all other physicalists, end up "reducing" qualia to physical states in a way that fails to do justice to their essential nature?

The answer to all of these questions is: no. The functionalist view, as typically stated, is indeed deficient in just the way critics of physicalism say it is, but at the same time, it gets at a fundamental truth about the nature of qualia which those critics, including proponents of the RIT, fail to see. The two sides can in fact be reconciled, and Hayek's position shows us how. And the reason why is summed up in the formulation I suggested earlier in this essay, and which I think Hayek is at least implicitly committed to by virtue of his adherence *both* to an indirect realist account of perception *and* to functionalism about qualia: *qualia are irreducibly subjective, but not intrinsic*. It is now time to see just what this means and how it provides us with as complete a solution to the qualia problem as we are likely to have.

(b) The inevitability of functionalism

Let us begin by noting some of the strengths of the functionalist position. As I said in the last chapter, what led so many physicalists to reject the (standard, materialist) identity theory in favor of functionalism was the evident *multiple realizability* of mental states, the fact that it appears at least possible for creatures of many different kinds of material composition to have minds. In effect, functionalism thus appeals to the same kinds of intuitions about conceivability that are usually used

---

between advocates of the two models in order to develop and defend those aspects of Hayek's view

to attack physicalism: whether or not one accepts that the mind can exist apart from *any* physical substance, it certainly seems clear that it can in principle exist apart from the one kind of physical substance we're used to seeing it in, i.e. the brain. The mind-brain identity theory thus appears illegitimately to limit the realm of the mental – and this is as true of the RIT as of any other version.

The other side of this coin is the question of why even the *brain itself* should be associated with mental phenomena, as opposed to things which obviously aren't, such as rocks, tables, trees, and the like. It is hard to see how this can be explained without some reference to the structure of the brain, the fact that it is put together in such a way that it is capable of supporting the sorts of functions associated with the mind. That is, it is hard to see how to explain this fact without referring to the *functional organization* of the brain, in particular the fact that it has a complexity of organization that rocks and the like lack. It can thus hardly be seriously maintained that functionalism has *nothing* of interest to tell us about the mind.[14]

Functionalism is also, in my view, supported by the fact, noted earlier, that qualia are describable only in terms of their relations, so that they *seem* to be ineffable:

---

which crucially depend on connectionist themes (and which we'll explore in the next chapter).

[14] Indeed, Russell himself seemed to recognize the importance of this sort of consideration, and we would do well to remember, as perhaps too many Russellians do not, that Russell not only said that our common sense conception of matter is in error, but also our common sense conception of *mind*: "The truth is, of course, that [the standard conceptions of] mind and matter are, alike, illusions" (1956, p. 135). Russell took modern psychology to have dramatically revised our view of mind in a way that parallels the revision of our view of matter wrought by physics; and he was especially impressed by the notion of the conditioned reflex, something only a system of a certain degree of complexity and functional organization is capable of, as being essential to a proper understanding of intelligence (p. 143-144). Most strikingly, he even says that an "event is not rendered either mental

this is precisely what one would expect if their character was entirely *determined* by their relations, so that there simply *is* nothing more to be said about them, which is precisely what functionalism claims. There are other reasons too to think that functionalism must at the very least be a part of any complete story of the mind's place in the natural world; and some of them are suggested by inadequacies in the RIT itself.

The Russellian view that it's qualia which flesh out the causal structure of the brain, though it arguably solves some puzzles, leaves us with others, for instance: why does the brain, from the "external" point of view, *look* so uniform if its intrinsic qualities are so radically different from one another? That is, given that its various states are identical with qualities as different as the qualia associated with looking at the sky, tasting a candy bar, hearing a symphony, feeling an itch, and smelling a skunk, why do brain states seem to observation so much alike? It is difficult to see how these questions can be answered except in terms of the functional role played by the brain states in question, i.e. by saying: the reason this brain state is associated with, indeed identical to, a red quale, is because of its relations to other brain states/qualia. Lockwood considers this move (which he attributes to Peter Smith) as a solution to the problem the question poses for the *traditional* identity theory; but it seems no less a problem for the RIT which he endorses. In any case, his reason for rejecting this solution is instructive:

---

or material by any intrinsic quality, but only by its causal relations" (p. 152)! So perhaps Russell himself was more a Hayekian, and less a Russellian, than might appear at first glance!

> [I]t is difficult to see how this view can constitute any advance over straight functionalism. For either functionalism has a problem accounting for intrinsic phenomenal content (qualia) in terms of functional role, or it doesn't. If it doesn't, then it is unnecessary to appeal to the intrinsic physical character of brain states at all. If it does, then the same problems will beset this mixed view, in regard to phenomenal qualities, as beset the simpler functionalist view. I have in mind, of course, its essential arbitrariness: the fact that it would seem possible, a priori, for any functional role to be associated with any phenomenal content, or with none. (1989, p. 126)

I think that Lockwood is entirely correct to say that if we accept this solution, we might as well go whole hog for functionalism and forget about appealing to the brain's intrinsic qualities at all. As I will be trying to show shortly, despite appearances to the contrary, he's *incorrect* in assuming that the latter isn't a good option; but suffice it for now to note that if we *don't* appeal to functional role, it's not at all clear what else we *could* appeal to.

Another, related, problem is this: if qualia are what flesh out the causal structure of the brain, why are some brain states and not others associated with qualia? Even if we assume, as Russellians generally do, that *all* parts of the material world are associated with qualia, so that all brain states are associated with qualia, this fails to solve the basic problem (in addition to leading to panpsychism, as we've seen). For the point is that it seems clear that not all brain states are associated with qualia *of*

which the subject is aware. Lockwood discusses the problem in terms of the analogy of an inner "searchlight," i.e. awareness, which scans the inner landscape of the brain, its various states – including qualia – becoming conscious when the searchlight hits them; and this "searchlight," he says, is identical to some neural process, the workings of which, he acknowledges, are mysterious (1989, pp. 163, 169). But however the process works, it is clear that it is the *relations* the various qualia bear to it that determine whether or not they are conscious. And thus we're led again to appeal to functional considerations as playing a part in explaining the nature of mental states, including qualia.

Lockwood would no doubt object that this assumes that a subject's being conscious of a quale is essential to it, so that if it must depend for being conscious on its relations, it cannot be an intrinsic quality; and this is an assumption he would deny, for Lockwood holds that qualia can exist unsensed by any subject. But this view, as we've seen, does not stand up to scrutiny. And even this reply isn't open to those Russellian sympathizers, such as Chalmers, who aren't committed to unsensed qualia. They might say, though, appealing to panpsychism, that brain states/qualia of which I am not aware are still sensed by *some* sort of subject or "proto-subject" and so are conscious, even though not related in the appropriate way to the sort of neural process Lockwood identifies with the "searchlight" of awareness.[15] Nevertheless – even if

[15] This notion of a proto-subject to which proto-qualia are presented is something suggested to me in conversation by David Chalmers. He also at least hints at this idea in his discussion of proto-qualia or proto-phenomenal properties at pp. 298-299 of his 1996.

panpsychism were acceptable, as we've seen it isn't – their being *conscious to the subject whose brain they are states of* does depend on the appropriate relations to this process, so that my point still applies. Nor would it seem to help to reply that it is *consciousness*, full stop, that is essential to qualia, not *consciousness to this or that subject*; for consciousness seems precisely to be consciousness to some subject or other, not something that can intelligibly be conceived of apart from a subject. Therefore if consciousness is essential to qualia, to change the subject to whom a quale is presented – from some "proto-subject," say, to me – would just be to change the quale, in which case the quale would not be intrinsic, but constituted by its relations, precisely as functionalism holds.

Further support for the suspicion that some sort of functionalism is inevitable is provided by Chalmers' arguments for what he calls the "principle of organizational invariance," which states that "given any system that has conscious experiences, then any system that has the same fine-grained functional organization will have qualitatively identical experiences" (1996, pp. 248-249). This, of course, is precisely what any functionalist would hold, but as Chalmers points out, someone who denied the functionalist claim that qualia *just are* functional states could also accept it. For all acceptance of the principle need commit one to is the idea that it is *empirically* impossible (even if not *logically* impossible) that something could have just the functional organization we do and yet lack just the sorts of qualia we have – in any case, this is all Chalmers wants to argue for, rejecting as he does functionalism (or,

more precisely, what he calls *reductive* functionalism, which he contrasts with *nonreductive* functionalism, which he accepts).

Chalmers' arguments try to demonstrate that even if absent qualia and inverted qualia are logically possible (which he accepts, which is why he rejects reductive functionalism), the assumption that they are *empirically* possible leads to absurdity; and since these are the standard scenarios used to show that functional organization and qualia can come apart, it follows that the assumption that their coming apart is empirically possible leads to absurdity. If absent qualia are possible, he says, then what he calls *fading* qualia are also possible; and if inverted qualia are possible, then what he calls *dancing* qualia are also possible. But it is absurd to suppose that fading qualia or dancing qualia are (empirically) possible. So it is absurd to suppose that absent qualia and inverted qualia are (empirically) possible.

The fading qualia scenario goes as follows. Imagine a system, which Chalmers calls Robot, having just the functional organization I do, but which is composed of silicon chips instead of neural tissue; and let's suppose for the reductio that this difference guarantees that this system, unlike me, is not conscious. Now imagine that my neurons are gradually replaced, one by one, with silicon chips, so that though my functional organization remains exactly the same throughout, my composition gradually approaches that of Robot until our cognitive systems are both composed entirely of silicon ships. Now by hypothesis, this means that at the end of the process, I am entirely lacking in consciousness. But at what point in the process did

consciousness disappear? One possibility is that at some point it simply blinked out. But this seems highly implausible: it amounts to the suggestion that the replacement of a single neuron would make all the difference between vibrant "technicolour phenomenology" (to use McGinn's expression, 1991, p. 1) and utter darkness, which would entail "brute discontinuities in the laws of nature quite unlike those we find anywhere else" (Chalmers 1996, p. 255).

The other possibility is that of fading qualia, of consciousness *gradually* fading out, qualia becoming gradually less intense until finally fading to black. But on this scenario, at the various intermediate stages, though my qualia gradually get more and more tepid – what once looked fire-engine red now looks faded pink, and so forth – my behavior is exactly the same (since my functional organization is the same), so I still say things like "Wow, does that fire engine look bright red – not at all tepid pink, that's for sure!" Moreover, if we construe belief in functionalist terms (which many would argue is plausible even if the absent and inverted qualia cases pose a problem for construing *qualia* in these terms), my beliefs will also be exactly the same as before: I'll *believe* the fire engine is bright red, even though the qualia it produces in me are faded pink! So in this case, we're stuck with a radical discontinuity between behavior, and perhaps cognition, on the one hand, and consciousness on the other; a discontinuity which, even if logically possible, would simply be unparalleled in the rest of the natural world, and so seems as empirically impossible as the idea of consciousness suddenly blinking out. But then, if fading qualia are empirically

impossible, the assumption that led us to assume otherwise must be false; that is, Robot, though of a different composition than me, must be conscious after all, sharing as it does my functional organization.

The dancing qualia scenario also involves a silicon duplicate with the same functional organization I have, but this time we are to suppose that Robot is conscious and has qualia, but qualia which are inverted relative to mine: where I see yellow, he sees what I would call blue, and so forth, even though our behavior, including our linguistic behavior (such as what things we *call* "yellow," "blue," etc.) is exactly the same. And let's suppose that in this case, too, there are a number of intermediate cases between me and Robot, where my qualia gradually approximate Robot's in qualitative character. Now take two stages with only a slight difference in qualia between them, stage A and stage B, the corresponding physical difference being a slight difference in the number of neurons replaced by silicon chips. Then suppose that at A I have a particular neural circuit supplemented with a silicon *backup* circuit, hooked up to a switch. When I flip the switch, the backup circuit takes over the work of the neural circuit, so that though my functional organization stays constant throughout, my composition suddenly switches to one just like that of stage B. It follows, then, that flipping the switch inverts my qualia, so that they become like those I'd have at stage B; and flipping it again, thus reactivating the neural circuit, reinverts them. Flipping the switch rapidly back and forth thus produces what Chalmers calls "dancing" qualia. But since my functional organization remains the same throughout,

my behavior, and arguably my beliefs (if we accept a functionalist account of belief), will stay the same: I will never say or do anything to indicate that my experiences are rapidly shifting back and forth, nor (arguably) will I even *notice* the difference! Again, we have a dissociation between behavior, and arguably cognition, on the one hand, and consciousness, on the other, and thus a discontinuity in the laws of nature, which is simply too radical to be empirically possible, even if logically possible. And if the dancing qualia scenario is empirically impossible, the supposition that led us to assume otherwise must be false; that is, inverted qualia must also be empirically impossible.

Chalmers' case for the principle of organizational invariance is bolstered, I think, by arguments like those of C.L. Hardin (1997) to the effect that our color space is asymmetric, so that it is arguably even conceptually impossible that the color spectrum could be inverted without some corresponding change in functional organization – even if it is conceivable that there be a creature with a symmetric color space and thus invertable color qualia.

Now Chalmers, again, does not endorse functionalism as usually understood, the sort he calls reductive functionalism because of its claim that qualia are not just associated with functional states in a law-like way, but are identical to such states. He remains a dualist in that he insists that qualia exist over and above functional states as well as over and above the material states which instantiate the latter. Nevertheless, I think it prima facie highly plausible to take Chalmers' (and Hardin's) results to be best accounted for by a thoroughgoing "reductive" functionalism of the sort Chalmers

remains wary of. For if we assume *only* a nonreductive functionalism of the sort he accepts, then we're left with a brute correspondence between functional organization and qualia that seems as implausible and theoretically unsatisfying as the Leibnizian parallelism of old, according to which mind and body are like two clocks keeping perfect time. On this sort of view, if we are to grant that the causal connection between consciousness on the one hand and cognition and behavior on the other is not merely apparent (as it is in the Leibnizian scheme), it seems we can account for it only by a baroque theory positing fundamental laws of nature tying qualia and functional organization together, irreducible in a way no other natural laws seem to be – a Rube Goldberg theory of consciousness. Nor does the RIT – with which Chalmers himself is a sympathizer – provide any way of avoiding this result. On a reductive functionalist theory, though, there is no mystery about the close connection between functional organization and qualia, no need to posit new, irreducible laws of nature: qualia *just are* functional states. So this sort of view, if it were otherwise unobjectionable, would surely be preferable to the alternative.

But precisely there, it will be thought, is the rub. However impressive is the case for functionalism, the view *does* seem at the end of the day to be objectionable, at least as a complete account of qualia. For the same objections I've argued are fatal against physicalism seem just as fatal against it: in short, there seems to be a conceptual gap between facts about functional organization and facts about qualia, in that it seems possible for a system to have just the sort of functional organization we

do, and yet have different qualia than we have, or even lack qualia altogether. So the most we can say for functionalism is that it is *part* of the story, but not the *whole* story – even if this leaves us, as with Chalmers' position, with a less elegant theory than we'd like.

I believe this is false, that a thoroughgoing *identification* of qualia and functional states *can* be defended despite appearances, and that Hayek's position suggests how. Before I say why though, let us take one more look at the standard case against functionalism – for on closer inspection, it is, I believe, less impressive than it appears, even without appealing to Hayekian considerations.

The best known argument purporting to show that functionalism cannot account for qualia is probably Block's (1978) "Chinese nation" argument, briefly discussed earlier. That argument, now to spell it out a little more thoroughly, goes as follows. It is conceivable that the entire population of China could be so organized that the system thereby constituted mimics in its functional organization the functional organization of the mind. We can suppose that a robot body with receptors mimicking our senses is hooked up by radio to the Chinese system in such a way that as its "senses" are impacted, signals are sent through the population in a way corresponding to the train of mental states following on a perceptual experience, and that these signals ultimately in turn generate behavior in a way corresponding to that in which mental states generated by perceptual experience do. Perhaps when the robot's receptors are triggered, a radio signal is sent to a bank of lights which signal part of

the Chinese population to send certain signals via walkie-talkie to other members of the population, then these other members send signals to yet other members, and so on until at last some final group radios signals back to the robot body which bring about behavior. Imagine that this complicated system of over a billion interacting individuals, serving as a kind of "brain" to the robot body, generates in that body behavior which is indistinguishable from that of an ordinary human being. What we'd have is a system that has exactly the same functional organization we do – the sensory inputs and behavioral outputs would be the same as ours and the intermediate links would parallel those of our minds in their relationships to the inputs, outputs, and to each other. But the system would clearly nevertheless lack mental properties, and in particular would lack the *qualia* associated with our mental states. Even if, when kicked in the shins, the robot would cry out and hop up and down cursing, it surely would not genuinely have a *sensation of pain*, the distinctive *quale* associated, in our case, with pain behavior. So functional organization and qualia can in principle come apart; in which case there must be more to the having of qualia than the having of certain kinds of functional states.

Now Block's thought experiment admittedly has a great deal of intuitive force; but that force is in my view largely stripped away from it another thought experiment reminiscent of those advanced by Chalmers. For reasons that will become obvious, I will refer to it as the "Spaghetti-head" scenario.

Imagine that neurosurgeons someday figure out a way to disentangle the billions of fibers constituting the brain in such a way that its functioning is not affected. Suppose a subject volunteers to live out the rest of his life on a table in a lab so that this can be done, his skull is opened up and his brain removed, and his neurons are slowly and carefully disentangled and hung from hooks above the table. We might also imagine that the neurons are treated with a new chemical which allows them to be stretched almost indefinitely without breaking or losing their conductivity. Eventually the room becomes filled with billions of tiny strands hanging from the ceiling (and arranged in any way the neurosurgeons see as conducive to realizing whatever ends for which they undertook this strange task in the first place – maybe they did it just to see if they could). All this time, though, the subject remains just as he was before the experiment, apart from the fact that he hasn't moved: he is as capable of having thoughts and experiences as he was before, and notices no differences in his mental life (other than, say an occasional queasiness brought on by the neurosurgeons tampering – or by the thought of what has happened to him). Obviously, this is all science-fiction of a sort not at all likely to be realizable. But it seems perfectly conceivable, and thus perfectly (logically) possible.

Now suppose that one day the neurosurgeons decide to carry their outlandish project a little bit further – actually, a *lot* further. Imagine they wheel our hapless subject, whom they've come affectionately to call "Spaghetti-head," outside the lab and onto a field they've prepared. Then they and their assistants begin stretching the

little fibers even farther, to the point where the field is now *covered* with fibers for miles and miles (again supposing, what seems, though empirically unlikely in the extreme, at least logically possible, that this can be done in a way that breaks no fibers and preserves conductivity in such a way that the signals from neuron to neuron are not significantly slowed). From the air it's quite a sight: a little body on a table, from the top of which protrudes an enormous, miles-wide, grayish cloud of threads. Still, Spaghetti-head, though understandably and increasingly ill at ease, retains his normal mental functioning.

Now imagine that the neurosurgeons start replacing Spaghetti-head's neurons in a manner like that described in Chalmers' thought experiments, except that instead of replacing them with silicon chips, they replace them with *people*. Specifically, when they remove a neuron, they attach a radio unit to each neuron with which it was connected, and give another radio unit to the person replacing it. Instead of sending a chemical signal, the neurons which previously triggered the replaced neuron send a radio signal which is picked up on the human replacement's radio, and he in turn sends further radio signals, in lieu of chemical ones, to other neurons just as the replaced neuron used to. Suppose at first a hundred or so neurons are replaced in this way. Just as in Chalmers' example, it seems highly implausible to suppose that mental functioning would be altered in any way; and in particular, it seems implausible to suppose that even the subject's qualia would be altered in any way. Spaghetti-head

would have exactly the sorts of qualia he had before, despite the change in his physical make-up, so long as his functional organization remains the same.

The reader has no doubt by now guessed where all this is going. By an extension of the thought experiment, we can imagine that *all* the neurons are replaced this way – perhaps with the entire population of China. And if it is implausible to suppose that the replacement of neurons with silicon chips in Chalmers' thought experiment would lead to any change in qualia, it seems no less implausible that Spaghetti-head – or China-head, as we might now wish to call him! – would undergo any change in *his* qualia. But then it is not at all *obvious* that Block's *Chinese nation* scenario would be an absent qualia situation after all. Nor does the "robot body" aspect of Block's thought experiment make it any more convincing: we can easily alter our thought experiment by imagining Spaghetti-head's body being gradually replaced with metal parts *before* his brain is tampered with, so that he becomes a brain attached to a robot body before becoming a *spaghetti*-brain attached to a robot body, before finally becoming – as in Block's example – a *China*-brain attached to a robot body. The upshot of the thought experiment seems much the same, with qualia plausibly being preserved throughout the various transitions.

What the Spaghetti-head thought experiment shows, in my view, is, at the very least, that the sorts of arguments typically given against functionalism are not nearly as conclusive as they appear at first glance. Indeed, I think it implies that those who appeal to outlandish scenarios to develop such arguments do not always think through

the details of the scenarios very thoroughly, but rely on merely impressionistic descriptions to generate the intuitions needed to make their case. Nor is a "counter thought experiment" like mine necessarily required in order to see this: I think if one really thinks through, carefully and in *detail*, *exactly* what would be involved in a Chinese nation type scenario, it's not at all clear that we wouldn't intuitively think of the system as having qualia. If one imagines a robot body reacting *exactly* as we do, tears and all, together perhaps with a hurt look when it is suggested that it is really only an unfeeling Chinese-nation-brained robot! – and imagines also what the "Chinese brain" would be like, what it would *look* like from the air, say, a complex, astoundingly orderly system, itself almost like an enormous living thing – it's hard to deny that one might at least wonder whether this thing might have qualia after all.

I think that the Spaghetti-head strategy can also be extended to serve as a counter to most other proposed anti-functionalist scenarios. For example, we can imagine that the various neuron-replacing individuals in our scenario each take off in rocket ships headed for different parts of the solar system (or even galaxy... universe?!) but equipped with radios that allow them to communicate with each other in just the firing-neuron-simulating way they did while on Earth. Spaghetti-head would thus go through his China-head phase only to pass from it to a more radical Solar system- (or galaxy- or universe-) head stage, holding on to his qualia all the while. And we can continue such extensions of the basic idea until, before long, we'll have reached a point at which the wildest, "sentient universe" speculations of science-

fiction (or of certain physicists while off duty, e.g. Frank Tipler, 1994) attain a certain dizzying plausibility. At the very least, they serve to undercut the "obviousness" of many suggested absent qualia scenarios.

Still, even if it is conceded that the Spaghetti-head type scenario takes the wind out of the sails of "Chinese nation"-style arguments, it might still be thought that this does nothing to touch the main point underlying all anti-functionalist arguments. Block's argument, it might be said, is only intended as a vivid illustration of something that can be known on independent grounds, namely that there is a *conceptual gap* between functional organization and qualia. If the spaghetti-head example is plausible, that's only because it inherits its plausibility from the connection we all recognize to exist between qualia and the *brain's* functional organization. But even *that* connection appears contingent. For the zombie and knowledge arguments show that there's no conceptual connection even between qualia and the functional organization of the brain. Nothing in the Spaghetti-head example does anything to show otherwise.

We reach, then, the nub of the matter. This response is, in my view, completely devastating against standard versions of functionalism, which are *physicalist* versions of functionalism. I do not believe those versions can have any convincing reply. But there is *another* possible version of functionalism, one suggested by Hayek's position – and it is not similarly helpless. It is time at last to see why not.

(c) Hayekian functionalism

Recall what Hayek has in common with the Russellian view – its commitment to the view that *all we're ever directly aware of are our own mental states* – and you are halfway to seeing what a Hayekian response to the opponent of functionalism is going to be. For Hayek, no less than for the Russellian, we are never directly aware of the external physical world, and lack any grasp of its intrinsic nature; at best, we have knowledge of its structure. And just as the Russellian argues that this limitation on our knowledge undermines standard objections to an identification of mental states and brain states, the Hayekian, I want to suggest, can argue that it undermines objections to an identification of mental states – especially qualia – with *functional* states.

The first thing to note is what precisely must be done in order to show that qualia can come apart from functional organization in the way anti-functionalist arguments allege. The zombie argument against physicalism in general requires a scenario in which all the physical facts are just as they are, and yet there are no qualia. More to the point, it requires the *conceivability* of such a scenario. And for such an argument to work against functionalism, what is needed is the conceivability of a scenario where all the facts about *functional organization* are just as they are, and yet there are no qualia. It might appear that the two tasks are really identical, or at least very similar: after all, if I've conceived of all the physical facts, aren't the facts about functional organization automatically included? Well perhaps so – *if* one is indeed

conceiving of *all* the physical facts. The trouble is that in going through the exercise of imagination or conception, one might easily believe he is conceiving of all the physical facts relevant to a domain, while a more careful consideration would reveal that he hasn't in fact been doing so after all. I might think that I have conceived of all the physical facts about the brain that there are when I think of a grayish, lumpy, wrinkled object, and perhaps even of the billions of neurons making it up; and I might therefore conclude that such an object just isn't capable of governing the complex behaviors exhibited by human beings. For it might seem to be not greatly different from a ball of twine – just with many more, and much thinner, "strings," packed more tightly. But of course, if I thought this, I'd be wrong: in conceiving this, I just *wouldn't* be conceiving *all* the physical and functional facts about the brain, because I wouldn't be conceiving of e.g. the relationships the different bits of neural "twine" or neurons have to one another in their dynamic aspects, while firing signals to one another at a staggering rate over the whole sweep of the brain, interior and exterior. When I conceive all *that*, it's much less hard to see how it can generate the sort of behavior associated with human beings.

Of course, none of this is meant to suggest that conceiving of more of *that* sort of thing would do the trick in the case of understanding the relationship between the brain's functional organization and qualia. As what I've said earlier suggests, I accept the anti-physicalist objection that further neurophysiological data is by itself incapable of removing the mystery. The point is just that it isn't always obvious that one is in

fact conceiving of what one takes oneself to be conceiving. And the problem here is exacerbated by the indirect realism and structuralism shared by Russellian and Hayekian alike: if we *don't* and *cannot* know what the physical world is like in itself, it's that much more difficult to be confident that one has successfully conceived all the physical facts, including all the facts about functional organization, relevant to a particular domain. Indeed, it's much harder to know that one has conceived anything but facts which are not *really* intrinsic physical facts at all. After all, in the case of imagining a grayish, lumpy, wrinkled object composed of billions of tiny fibers, and even of those fibers as furiously sending electrochemical signals to one another, I haven't really imagined the intrinsic nature of the brain – I've only imagined a number of qualia of the sort typically caused by brains. If I try to think of the brain itself, the best I can do is to try to imagine something very abstract – a certain kind of system of such-and-such a causal structure. And in doing so, there is always likely to slip into my conception some feature that strictly shouldn't be there. Indeed, I ought really to try to stop *visualizing* the brain at all, since visual concepts are almost completely tainted in that they involve features that are not qualities of the physical world as it is in itself. It is almost like trying to imagine what a number is like (if it's "like" anything!) or to do mathematics in one's head without thinking of the concrete symbols we use to write out mathematical formulae. It would be foolish to suppose that the mathematical facts themselves are in their intrinsic nature anything like mathematical symbols. We need the symbols to think effectively about mathematics at

all, of course; but if we're trying to think of what the mathematical realm is like "in itself," we have, of course, to try to abstract away from them. And by the same token, even if it is extremely difficult to think of physical objects in any way other than in terms of the sensible qualities by which they are represented in our experience, if we want to attempt anything like a conception of such objects as they are in themselves, we have to abstract away such qualities.

I believe that even those sympathetic to the Russellian point of view fail to realize the implications it has for the thought experiments they use against functionalism – the very same sorts of arguments they typically reject when applied to (the Russellian version of) the mind-brain identity theory! To take seriously the limitations on our knowledge the Russellian rightly stresses is, I want to suggest, to see that it is not at all clear that those thought experiments have the force they seem to have on a naïve conception of matter. For when we abstract away everything that indirect realism and structuralism require us to abstract away, it simply is not at all obvious that there's enough left which can be said with confidence to be able to come apart from qualia. Note the way Chalmers, for example, defends the zombie scenario by appealing to the similarity in functional organization between the human brain and the silicon brain of a robot or a Chinese nation type system: "For it is clear that there is no more of a *conceptual* entailment from biochemistry to consciousness than there is from silicon or from a group of homunculi" (1996, p. 97, emphasis in the original). His point is that if absent qualia are possible in e.g. the Chinese nation case, they're no

less possible in the case of the – functionally isomorphic – brain. But notice that he doesn't say there is a lack of conceptual entailment from *functional organization* to consciousness, though that's no doubt part of what he means to imply; he says rather that there's a lack of such entailment from *biochemistry*, and *silicon*, and *homunculi* to consciousness. This is perfectly understandable given that it's very hard to imagine functional organization *apart* from its realization in some such physical system. But at the same time, as the Russellian view has taught us, the exercises in conception that we typically undertake when we conceive of biochemistry, silicon, and homunculi are not, after all, cases of conceiving these things as they are in themselves. *Really* to conceive of them in a way that strips away all of the misleading perceptual features would in fact be to imagine *nothing but* causal structure. And if we go about zombie type thought experiments *that* way, it's much less clear what the results are – for it's much less clear what exactly we're conceiving. The force of the thought experiments in question, I want to suggest, actually depends surreptitiously on an implicit assumption of just what the Russellian in his more thoughtful moments would deny himself, namely a conception of physical systems in terms of the sensible properties through which we are acquainted with them in perception. "Surely qualia can come apart from functional organization," even Russellians assure us, "since there's no conceptual entailment from biochemistry, silicon and the like to qualia." But what do biochemistry, silicon, and the like *conceived of in terms of their sensible properties*, have to do with functional organization? The lack of a conceptual entailment only

seems *obvious* so long as we have failed completely to free ourselves from the naïve conception of matter the Russellian is otherwise so concerned to undermine.

But now consider something that Russellians seem not to realize but which also follows from the indirect realism and structuralism they, along with Hayek, endorse, and which shows that our grasp of the structure of the physical world is more tenuous than even what I've said so far suggests. Perception, it is agreed, fails to reveal to me what physical objects are like in themselves. But it's not as if, in perceiving a particular physical object or objects, I *am* (in some sense or other) directly aware of its *causal structure*, though not of its intrinsic properties. I'm not *directly* aware of *it*, or of *any aspect* of it, *at all*. I'm directly aware, instead, of certain mental phenomena. And it's only by positing an external object having such and such a structure as the external cause of my perceptions, as part of a kind of explanation of those perceptions, that I have any knowledge of the external objects *and their structure* at all. In other words, all of what I know about external objects, *including their structure and functional organization*, is *theoretical* knowledge. It isn't of the nature of *data* which must be accepted by any party to a dispute about the nature of external objects. And it is possibly an implicit assumption that things are otherwise that adds to the sense that the anti-functionalist thought experiments have intuitive force even from the Russellian point of view.

The *next* thing to note is that on the Hayekian point of view, in introspecting our own qualia – in introspecting the "sensory order" – we are directly aware of a

domain the elements of which we know only in terms of their relations to one another, so that what we know of *it* is only structure. And this is, I believe, not even prima facie as implausible a view to take as it might seem. For as has already been noted, part of the problem of describing qualia in the first place is that it seems impossible to do so *except* in terms of their relations. Surely that provides at least some defeasible support to a structuralist, indeed broadly *functionalist*, view of qualia. For it is just what we should expect if qualia are indeed features whose nature is determined entirely by their relations. (Their nature *as qualia*, that is: obviously whatever it is that plays the role defined by a quale's relations has *some* intrinsic nature. The claim under consideration here is just that that intrinsic nature is not what is revealed to us in introspection, and is in fact unknowable, so that our knowledge of qualia is not knowledge of intrinsic properties as such.)[16]

Let us take stock. We have only the most tenuous conception of the functional organization or structure of the material world, including the brain, when we abstract away from what perception reveals to us, as we must on the Russellian view. *All* of what we know is theoretical, that is, it is a matter of being justified in believing propositions the constituent concepts of which have somehow to be derived from what we do know directly, without inference. But what we know in this way is just the

[16] I would (tentatively) suggest also the following positive argument *against* intrinsicality: As Strawson (forthcoming) suggests, causal interaction is arguably sufficient for same substancehood. But then, since the brain and objects in the external world causally interact, they must be of the same substance. Now as we've seen (in our discussion of panpsychism in the last chapter), qualia can't be the intrinsic properties of objects in the external world. But then, since the brain is of the same

sensory order of our own qualia; and in introspection of that order, it is at least plausible that we have a direct grasp of something that we nevertheless know only in terms of its relations, its structure.

The implication should be obvious: if the Hayekian is right, it is *precisely in our conception of qualia, derived from introspection, that we have our clearest grasp of functional organization or causal structure. And our conception of the functional organization of anything else – including anything outside the mind – must be, or at least (plausibly) is, derived from the conception we have via our awareness of the relations existing between qualia.*

The answer to the anti-functionalist I am suggesting is this, then: the standard arguments against functionalism about qualia fail for the same reasons the Russellian takes them to fail when applied to the RIT, namely that they assume a conception of the external material world to which they are not entitled. In particular, anti-functionalist arguments assume that we have a conception of functional organization on which it is clear that the functional organization of the brain could come apart from qualia. But not only do we *not* have such a conception; the best candidate for a conception of functional organization we do have is precisely the realm of qualia itself. And this realm is something that there are independent (if defeasible) reasons to identify with the brain, or, more precisely, with the brain *qua* what instantiates a causal structure isomorphic with that of the mind. So there is good reason to identify qualia

---

substance as those objects, they can't be the intrinsic properties of the brain either. But they are

with functional states (namely the case for functionalism made above), while the standard objections to such an identification are undermined by the fact that they presuppose a conception of functional organization to which they are not entitled. At the very least, those objections can be rejected as question-begging: for since we have no conception of functional organization except for that derived in some way from our introspection of qualia, it seems the anti-functionalist can have no *independent* grounds for claiming that qualia can come apart from functional organization.[17]

The Hayekian response to such arguments as the zombie argument and the knowledge argument is thus obvious: to conceive of a system having a functional organization exactly like ours would *just be* to conceive of qualia like ours; so for a world to have just the functional organization ours does would just be for it to contain qualia, and not be a zombie world at all, and for someone like Mary to have *all* the knowledge there is to have about the brain's functional organization would just be to know "what it's like" to have qualia.[18]

---

properties of the brain; so they must be *non*-intrinsic properties of the brain.

[17] To get an absent qualia or inverted qualia argument against the Hayekian view going, then, one would first need to establish that a functional organization like ours can be conceived of in a way other than the way the Hayekian conceives of it; but then such an argument would be otiose, for to establish this would just be to establish that that functional organization can come apart from qualia.

[18] Of course, since scientific knowledge plausibly includes knowledge of functional organization, this might entail – contrary to the presupposition apparently shared by both Jackson and his opponents – that there is some scientific knowledge that cannot be had by those lacking a given sensory modality, so that the congenitally blind, for example, cannot after all have a *complete* knowledge of the neurophysiology of vision. But why should this be surprising? After all, someone lacking *any* sensory modalities wouldn't be capable of any scientific knowledge at all; so it isn't at all odd that those lacking a particular modality should also lack a particular kind of scientific knowledge. As Hayek says (in a paper he should perhaps have titled, following Nagel, "What is it Like to Be a Rat?"), in order *fully* to understand *all* the facts about a rat, "we would in effect have to become another rat" (1982, p. 293). (That this is impossible indicates that Nagel's problem, which is related

The meaning of the formulation I suggested earlier should now be clear. Qualia are *irreducibly subjective* in that they are what we are directly aware of in the first-person realm of perception, and are not accessible from the third-person point of view. The view I am defending is thus in no way committed to explaining away, or defining away, or in any other way surreptitiously *doing* away, with what is essential to qualia – in the way all physicalistic theories, with their insistence on reducibility to the third-person point of view, seem to be. At the same time, qualia are *not intrinsic*: the nature of a quale, *as a quale*, is entirely determined by its relations (but relations to which we have direct access only from the first-person point of view). My view is thus a kind of functionalism – though not a physicalistic version of functionalism.[19] And while Hayek was himself apparently not *explicitly* committed to this view, it was inspired by his work and I believe he would have accepted it, so that I think it appropriate to label it *Hayekian functionalism* or HF for short. It might be thought that it too is a kind of "nonreductive" functionalism, since it doesn't reduce qualia to third-person physical features. But I'll cede to Chalmers all rights to that label, partly to avoid confusion but partly because HF might indeed plausibly be regarded as a reductionism of sorts. Only it's not quite the case that qualia get "reduced" to

---

to Jackson's problem, cannot *fully* be solved. I'll say more about the limits Hayek thinks there are to our knowledge of the mind, including to some extent even our knowledge of qualia, in the next chapter.)

[19] Strawson seems sympathetic to something close to this view – though, to be sure, not to this view itself – when he endorses a qualified version of D.M. Armstrong's thesis that "a mental state is, essentially, 'a state that is *apt to be the cause of certain effects, and apt to be the effect of certain causes*'" (1994, p. 260, emphasis in the original), where these effects (contra Armstrong) do not

functional states on this theory; at least, it's misleading to put things this way. For given that it's the sensory order itself that gives us our best conceptual grasp of functional organization, we might say more accurately that functional states get "reduced" to qualia. That is, just as on the RIT, it's not that qualia turn out to be brain states so much as that brain states turn out to be qualia, on HF it's not that qualia turn out to be functional states so much as that functional states turn out to be qualia! We might say that Hayek does for functionalism what Russell (and others) do for the identity theory. "Hayekian functionalism" is thus an apt counterpart to the "Russellian identity theory."

Not only is HF immune, in my view, to the criticisms typically made to other kinds of functionalism, and thus inherits all their assets without any of their liabilities; it also has unique advantages of its own. For instance, I believe it accounts for the fact that mental states *seem* so very different from the brain even though they're identical. Of course, the indirect realism and structuralism it shares with the Russellian view is part of the reason, but there is more to it than this. Given the nature of the brain's classificatory activity in perception, of grouping objects and processes according to their similarities with respect to certain features, it's no surprise that the brain, even if identical with the mind, should *seem* so similar to other objects of perception: *anything* the brain classifies is going to be grouped with other objects of classification, in some respects and not others – including itself and its own processes. So *of course* it's

---

essentially include bodily behavior, and where (contra Armstrong and the view I'm defending) the

going to seem like just one other physical object among others so long as one is aware of it via the same process, perception, by which it is aware of other external objects.

It follows that there should be no surprise that the first-person, subjective, phenomenal world seems so different from the third-person, objective, external world. For suppose brain state B is the one produced by light from a ripe tomato hitting your retinas, etc., and is thus, on this view, identical with your reddish, tomato-ish quale. Now you open up your skull and in a mirror see the part of your brain wherein B obtains. "It seems so different from my quale!" Well, what did you expect, that you'd see a reddish, round-ish object in your brain? After all, what you're directly observing *now* is not B anymore, but *another* brain state, call it C, which is caused by light from B striking your retinas, etc. You *never directly* get at your brain states through *perception*, after all, but only through *introspection*; in perception, you only get directly at other brain states which represent them. And so on for your perceptions of those *second-order* brain states. In fact, introspection itself can be seen as (and it's clear from The Sensory Order that Hayek himself saw it as) just an extension of this process: perception groups external objects in terms of their relations, introspection groups internal objects in terms of *their* relations. So perhaps, in a sense, we never *directly* are aware of anything; awareness of the physical world is mediated by qualia;

---

character of the states is not *entirely* determined by the causal relations involved.

awareness of qualia by higher-order mental processes; awareness of those by yet higher-level processes; etc. etc.[20]

This also explains *why* our knowledge of the external world is knowledge only of structure. If perception is just a matter of classification of external objects according to similarities and dissimilarities, then we should expect that all we can know of those objects is their *relations* to one another. And insofar as we have (independent of the Hayekian view) evidence that the brain acts in perception and introspection as a classificatory system, this lends support to the Hayekian thesis that even knowledge of the sensory order of qualia is knowledge only of structure: if the brain, in introspection of perceptual states, just repeats at a higher-level the sort of classificatory process that occurs in perception, grouping those perceptual states according to similarities and dissimilarities, then we should *expect* that all the knowledge we have of the internal world should also be knowledge only of relations, of structure. (Indeed, if *all* knowledge somehow can, by extension, be seen as a further iteration of such a classificatory process, it would follow that we can never, in the nature of the case, have knowledge of anything but structure, can never have knowledge of intrinsic qualities. I will not try to argue for this bold suggestion here, though.)

---

[20] As Lockwood notes (1989, p. 169), Kant also thought that our knowledge even of the internal world of the mind was "mediated by a veil of representations," even though Lockwood also counts Kant as a precursor of the Russellian view. This aspect of Kant's position, together with (to speak anachronistically) the Russellian elements, I think makes Kant more a precursor to Hayek than to Russell – there being other Kantian themes in Hayek in any case, as we've seen.

It is really this "higher-order" character of perception, introspection, and perhaps knowledge in general that gives rise to a subjective realm at all, and makes the mind as mysterious as it is – not only in its phenomenal, qualia-related aspects, but in other respects as well, as we'll see in the next chapter. And the recognition of the fact that we are ever only *indirectly* aware of the external world provides the key to seeing *how* exactly this is so. That they ignore this fact is, in my view, why physicalistic-minded functionalist theorists who otherwise see that there is some connection between the mystery of consciousness and the notion of higher-order mental processes (e.g. Hofstadter, 1979, Chapter XX; 1981) never quite tie them together in a way that seems to *solve* the mystery in a satisfying way, so that there doesn't seem to be some unexplained residue left over. Such theorists never escape the naïve conception of matter which physicalists accept along with common sense and to which indirect realism and structuralism show we are not entitled. Thus they characterize complex systems like the brain, higher-order processes and all, in exclusively third-person terms.[21] And so they not only leave out the – distinctively first-person – realm of qualia, but also arguably distort the character of higher-order mental processes: if we *start* with a recognition of the mind's always *indirect* contact with reality, and add from our understanding of the brain the notion of classification, the idea that the peculiar character of the qualia through which we perceive the world is a result of their

[21] In short, the problem with materialism is that it starts from "outside" the mind and tries to work its way back in; while the proper approach is to start from *inside* the mind (the only place from which we can start anyway) and work outward to the external world.

higher-order classificatory character (since what is classified will always, by the very nature of the process, seem to be of a fundamentally different sort from what does the classifying) falls out quite naturally. The subjectivity insisted on by critics of materialism comes to seem less mysterious and occult, and more to be precisely what we should expect in a system that interacts with its environment via the having of internal states which are classified by other internal states, which are classified by yet other ones, and so on.[22]

At the same time, the commitment to indirect realism ensures that the first-person realm of subjectivity is not "reduced" to some material process understood in physicalistic terms: again, we might say that it's not that subjectivity is reduced to (physicalistically understood) higher-order states so much as that higher-order states are "reduced" to subjective, phenomenal, qualitative ones. This should dispel any lingering suspicion that HF, in characterizing qualia as relational properties, really does do as little justice to them as standard physicalistic functionalism does.[23] The objection here would be that if someone takes qualia seriously at all as a difficulty for physicalism, he does so precisely because they are thought *essentially* to be intrinsic qualities, so that HF is really, in effect, as deflationary a view as Dennett's view that there simply *are no such things* as qualia, or that to the extent that there are, they are nothing more than an organism's capacities for differentiating or discriminating

[22] As before, "classification" in this context is to be understood in the non-intentional sense. We'll discuss the problem of intentionality itself in the next chapter.

between features of its environment (and indeed, this sounds very close to Hayek's own way of formulating his position, speaking as he does throughout The Sensory Order of perception as a process of classification). But as I've said, I would deny that what is essential to qualia, and what is *crucially* left out of physicalist accounts, *is* intrinsicality. Rather, what is essential and what physicalism can't accommodate is, I want to suggest, what has been variously described as the qualitative character or "feel" of qualia, that which makes for there being "something it is like" to have a quale, or what I've been calling their subjectivity or "first-person" character (and which, as I've argued already, entails also their direct apprehensibility and privacy). And the Hayekian functionalism I'm advocating preserves all of this whole.

What physicalism in all its forms does is force everything into the Procrustean bed of "third-person" accessibility, so that it is impossible to see how a person, if all he is is a material system (understood now, not in structuralist, but in "commonsense" materialist terms) can be any more conscious than a rock. On HF, though, all we're ever aware of, directly, is the "first-person" world; and as I've suggested, our awareness of that world is our *best source* for a conception of a functional organization like that of the brain, indeed, it is awareness of a world which *just is* a continuous stream of functional states, of classificatory activity. So to conceive of something having just the sort of functional organization I do would *just be* to

---

[23] I thank David Chalmers for voicing this suspicion and forcing me, in discussion of the issue, to clarify more precisely the respects in which Hayekian functionalism differs from standard functionalism.

conceive, not a zombie-like, Dennett-ish, qualia-free machine, thought of in naïve commonsense materialistic terms (i.e. as, in effect, nothing but a complicated rock) *but a system with all the qualitative feel, subjectivity, etc. that I have*. This sort of "guaranteed conceivability," if you will, is precisely what physicalism fails to give us, which is what makes it the case that it "leaves out" qualia; and "intrinsicality" or the lack of it is irrelevant to such guaranteed conceivability. So HF doesn't leave out anything essential to qualia or "deflate" them in any way; on the contrary, what it does is rather to *inflate* or "beef up" our notion of a functional state by making qualitative, first-person character essential to it. It's not that the mind turns out to be a colorless mechanism (as it seems to on standard functionalism), so much as that (complex enough) mechanisms turn out to be "technicolour phenomenology"-having minds (to use McGinn's (1991, p. 1) expression).

Now in fact intrinsicality, of a sort, thereby does come into play here; for if I'm right, then there being "something it's like to be" in a functional state is an intrinsic feature of a functional state, in that it is essential to a functional state. But that's only because what a functional state *just is*, on HF, is the sort of state that has a "qualitative character" or feel to it, etc. This doesn't contradict anything said earlier, for *qualia* still aren't intrinsic qualities, much less the intrinsic qualities of the *brain*. We might put things this way: there is, intrinsically, something it's like to be in a functional state or quale – however such states are physically realized – but nothing is by itself *intrinsically* a quale (it's only one when it plays a certain functional role) and the *way*

it is a quale in any individual case, the *particular* qualitative character a given functional state or quale has, is entirely determined by its relations to other functional states or qualia. In my awareness of, say, a red afterimage, I'm aware of a brain state alright, but not of any of its intrinsic qualities. Rather, I'm aware of qualities it has only by virtue of its being a functional state, and moreover, only by virtue of its being a functional state of a particular sort; though by virtue of this, I'm also aware of the intrinsic nature of *that sort of functional state.* (Compare: In being aware of me, you would be aware of a husband, even though being a husband is not an intrinsic property of me, but only a property I have by virtue of being married; though in being aware of me as a husband, you *would* be aware of the intrinsic nature of a state I'm in, i.e. the state of being a married man.)

Now all this suggests another – perhaps surprising – way in which HF has advantages other views do not: HF affords a complete solution to *the problem of other minds.*

That problem is just the problem of explaining how one can possibly be justified in believing that anyone other than oneself has thoughts, experiences, and mental states in general, given the gap that seems to exist between knowledge of another person's behavior, physiology, and the like on the one hand, and knowledge of his mental states on the other. For it is logically possible that a person could be exhibiting just the behavior, and have just the neurophysiological states, which we associate with particular mental states, and yet lack those mental states altogether. So

how can knowledge of the behavior and neurophysiological facts – which is all we have to go on – possibly justify us in claiming to know about other people's mental states? It seems that we cannot even know that it's *probable* that other people have such states. Such knowledge would have to be arrived at by observance of an appropriately large number of correlations between behavior, say, and mental states. But there is only *one* case in which such a correlation can be observed, namely one's own; and observation of a single case is obviously an extremely slender basis on which to conclude that a general correlation exists.

Now the gap between neurophysiolgy, behavior, and the like and mental states is precisely the gap illustrated in the zombie and knowledge arguments: the problem of other minds is thus linked in an intimate way with the mind-body problem.[24] And so no theory on which there remains a gap between neurophysiological and behavioral facts and facts about the mind – as there is even on the RIT, as I've argued – can solve the other minds problem. But suppose that HF is correct and that qualia are identical to functional states, so that we know what it's like to instantiate functional states of the sort we instantiate simply by virtue of our introspective knowledge of our own qualia. Then it's easy, in principle, to know whether or not anyone else has experiences of the sort we have: all we need to determine is whether or not he has a functional organization like ours. And since everyone can know that other human

[24] It is surprising that this is not more commonly noticed, though Nagel (1974, n. 14) notices it and says – rightly, as I'll try to show – that if we could solve the latter problem, we would automatically be able to solve the former.

beings, at least, have the same functional organization that he has, it follows that he can know that minds exist other than his own.

Again, on other theories, even the RIT, no such solution is possible. Perhaps I can know that *my* brain states are identical to qualia – that is, that qualia are what flesh out the causal structure of my brain. But how do I know that this is true of anyone else? After all, all I know about other people is their causal structure. On HF, to be sure, this would be enough, since part of that causal structure would be identical to the functional organization that HF says can't exist without qualia. But on the RIT, qualia are intrinsic properties distinct from any functional properties of a nervous system. And it follows that it is conceivable that something other than qualia might be what fleshes out the causal structure of any other person's brain. Of course, Russellians do go on to say that qualia are what flesh out the causal structure of the entire material world; and they might respond that, insofar as there is a problem about determining whether other people's brain states are identical to qualia, it is just a part of the larger problem of whether *any* part of the external material world is fleshed out by qualia, in which case the whole problem is just a version of general skepticism about the external world, and thus not a *special* problem for the RIT as such. But the point is that the RIT thus has no better resources for dealing with the problem of other minds than any other theory does, while HF does have unique resources; so that, if ability to solve more problems is a criterion of theory choice (and it is), it follows that, on this score at

least, HF is preferable to the RIT and any other theory which is unable to solve those problems.[25]

A final advantage HF has over other theories, including the Russellian view – which, as what I've said indicates, I take to be its only plausible rival – is that HF is able to solve the "grain problem" (Foster, 1991, pp. 126-130; Lockwood, 1992), the problem of explaining how it can be that qualia seem so much "smoother" and less finely grained in structure than brain states do if they are *identical* to brain states. Even if the Russellian combinatorial sort of explanation of the phenomenal character of a quale we considered earlier didn't have the problems it does have, it seems that the grain problem would remain; for a quale just doesn't appear to have the fine-grained structure that the combinatorial sort of explanation appeals to. From the point of view of HF, the RIT simply looks in the wrong place to account for a quale's phenomenal character. It's not the micro-structure of a particular brain state, but rather a brain state's relations to other brain states, that gives it its phenomenal character. More to the point, it's not a brain state as such that a given quale is identical with in the first place, but a functional state. Anything that played the functional role in question would have the same phenomenal character, whether it was something as complex in structure as a brain state, or something *more* or *less* complicated in structure. So as far as "grain" is concerned, what we need to worry

[25] Of course, the solution just suggested would also be open to standard, physicalistic functionalism, since it too identifies qualia with functional states; but unlike HF, it can be accused of leaving out precisely what is essential to at least conscious mental states, so that all things considered its ability to solve the other minds problem is not enough to recommend acceptance of it.

about when identifying qualia with brain states is not the complexity of the structure of a quale versus that of the structure of a brain state, but rather the complexity of the relations between qualia versus that of the relations between brain states. And it seems fairly uncontroversial that the two sets of relations *are* similar in complexity. That is, even those who doubt that qualia can be *reduced* to functional states don't seriously doubt that there is a discoverable isomorphism between the relationships holding between qualia, on the one hand, and those holding between brain states (at some relevant level of description), on the other. Once again, then, we have a case where HF is superior to other views, including the Russellian view – in this case, in its ability to deal with a difficulty which even Russellians acknowledge is the most serious problem facing the RIT (Lockwood, 1989, pp. 16, 177; 1992).

The advantages I've claimed its non-physicalistic identification of functional states with subjective, first-person ones gives HF might, however, appear to be offset by a disadvantage deriving from the same source: given its characterization of functional states, it might be suggested that HF surely entails panpsychism no less than the RIT does. The idea here would be that if having a functional organization like ours is sufficient for having qualia, and moreover, if (all) functional states by their nature have a certain qualitative feel to them, it would seem to follow that all sorts of things can and do have qualia – not just robots and computers, but *anything* with any level of functional organization at all, which seems pretty much to include nearly everything in the universe!

I think it must be admitted that if HF (or any other kind of functionalism, really) is true, it would follow that it is in principle possible for many things other than human beings and animals to have qualia – not just robots and sophisticated computers, which would more or less approximate human and animal motor and/or cognitive functioning anyway (so that the ascription of qualia would perhaps seem less jarring in these cases), but other, less anthropomorphic systems too. Still, this doesn't open the floodgates to panpsychism, nor, to the extent it does, like the RIT, entail the existence of qualia in surprising places, does it face the problems the Russellian view does on that score.

Even if HF, like the RIT, had to posit qualia or proto-qualia all the way down to the level of fundamental particles, and thus was as panpsychist as the RIT, it would still be preferable to the Russellian view on grounds of simplicity: the RIT would have to posit two sorts of thing, functional states and qualia, the latter associated with the former all the way down, given the principle of organizational invariance, while HF need posit only one, since it identifies functional states and qualia.

In any case, HF *doesn't* need to posit qualia all the way down. The RIT must do so, since it makes qualia or proto-qualia out to be the intrinsic features of the universe, so that they are at the bottom of things *whatever* a thing's functional organization or lack thereof. But HF need at most posit qualia only where we must assume there to be the instantiation of functional states. If functional organization

fades to zero, so that we get to basic elements having no functional organization at all, we needn't posit qualia or proto-qualia at that level at all.

Moreover, how precisely to conceive of the gradation from qualia down, through ever simpler levels, to proto-qualia is, as we've seen, problematic on the RIT, since it proceeds by thinking in terms of combinations of fundamental elements (on analogy with atomistic models in physics). But HF's identification of qualia and functional states gives us some traction in conceiving of simpler kinds of qualia than we might be familiar with in introspection, since we can proceed by focussing on the idea of ever simpler levels of functional organization. To see how, consider Hayek's own way of spelling out what sort of functional organization underlies a given quale, which is to speak in terms of impulses within different sets of neural connections being initiated by different aspects of a given stimulus, some sets of connections associated with some properties, others with others. If we start with an example like that of the quale involved when one looks at an orange, then, we can begin to conceive of simpler qualia simply by abstracting away the functional elements (neural impulses, say) associated with some of its features (roundness, say).

This leads to a further point: it is implausible to suppose that HF will, in fact, lead to too liberal an ascription of qualia. For if anything close to Hayek's way of spelling things out is correct, it is clear that even so simple a quale as that involved in seeing an orange must involve a very high degree of functional organization. Abstracting away the various functional elements soon leaves very little that can be

regarded as a quale at all close to those we are familiar with in everyday experience. So there is no danger of HF leading to such counterintuitive results as the ascription of sensations of hot and cold to thermostats, or whatever. Moreover, it might also be argued that even before functional organization grades off, qualia disappear from the scale – just as life surely does when we are considering complexity of organization. There is no clear dividing line between life and non-life, but that doesn't keep us from being sure that bacteria are alive and weather systems are not, with the case of viruses being a toss-up. Similarly, there is arguably no difficulty in claiming, say, that all organisms capable of making discriminations between elements of their environments (at one end of the spectrum) have qualia, and fundamental particles (at the other end) do not, with some intermediate systems being problem cases. (Presumably some of the problem cases could be decided by appealing to the fact, alluded to earlier, that qualia seem to require a point of view.)

In any case, by the time we get down to the most rudimentary of proto-qualia (wherever that is), they will clearly be so rudimentary that it would perhaps be misleading to think of them as *mental* properties at all, just as it would be to think of a car as being alive. That is, panpsychism only seems implausible to the extent we are thinking of mentality on the model of our own minds, with all their complexity. If we think instead of very rudimentary things as having at best something remotely like one particular feature of our minds (just as cars and the like have features remotely like some features of living things), then HF can even less plausibly be thought to commit

us to panpsychism as that view is usually thought of (though admittedly, this last point applies to the RIT as well).

The bold claims I have made for Hayekian functionalism thus appear to withstand scrutiny: it has all of the advantages of its rivals, and none of their disadvantages. I conclude that it is the best solution available for the problem with which we've been concerned, the qualia problem; and I think I've also shown that it is a solution which is *satisfying* in a way other alleged solutions are not, in that it removes the air of mystery surrounding the relationship between mind and matter.[26] Or in any case, it does so to the greatest extent possible. For there are grounds, also explored by Hayek, for thinking that we can never, in the nature of the case, *completely* understand the mind. This suggestion shall be our focus in the next and last chapter of this study.

[26] That it (allegedly) removed this mystery was in Russell's view the chief thing to be said in favor of his own position (1956, p. 153). If I am right, the Hayekian view has a more plausible claim to this distinction than Russell's does.

7. Hayek and the limits of knowledge

(a) The inscrutability of mind: consciousness, intentionality, and rationality

Our knowledge of the sensory order of the mind as well as of the physical order is knowledge only of structure, not of intrinsic qualities. The sensory order is identical to that specific *part* of the physical order whose structure it mirrors, namely the *neural* order existing in the brain, and our knowledge of the wider, external physical order is mediated by our knowledge of the mind/brain. All we know directly are states of the latter order, and all we know directly of *them* are the relations existing between them, the relations revealed to us in introspection of the constellation of qualia which constitutes conscious experience. In particular, all we are aware of directly are the *classificatory* states of the brain. We are aware of external objects and events only via our direct awareness of internal states which classify those external objects and events according to their relations to other such objects and events. Those internal states derive their own character as qualia from their relations with other internal states – that is, in introspecting *them* (as we do when our mental "focus" turns from the external object we take ourselves to be perceiving – a tomato, say – to the qualia through which we perceive it – a reddish patch, for instance), we do so only indirectly, via *further* classificatory states which classify qualia according to their relations to each other.

This, at any rate, is the picture painted by the view I've called Hayekian functionalism, and I've suggested that it provides as complete a solution to the qualia

problem as we're likely to have. For given that we lack knowledge of anything but the structure of the sensory and physical orders, there is no barrier to the identification of the former order with part of the latter – especially since even our most direct knowledge of the nature of causal structure *itself* is knowledge of the sensory order of qualia. We simply lack any knowledge of the physical world or even of its functional organization which would undermine the claim that the world of qualia, characterized in all its subjective, first-person glory, is identical with a part of it. The evidence that physicalists appeal to to support an identification of the mind with the brain, or at least with the brain *qua* what realizes a certain kind of functional organization, can thus be reconciled with what we know about the first-person, subjective, private character of qualia. Furthermore, the *classificatory* nature of subjective states helps to account for why mental phenomena *seem* to be so different from physical phenomena: the *perception* of external objects constitutes one level of classification, while the *introspection* of the internal states which classify those objects is itself a *higher-order* level of classification, so that the categories into which the respective objects of knowledge are put in the two cases are necessarily of a different kind – but the difference is (for all we know) thereby a difference *only* in the method of classification, as determined by the operation of the classificatory system, *not* (necessarily) a difference in the intrinsic nature of the thing classified.[1]

---

[1] An extension of this sort of explanation might account for why *thoughts* and the like also seem so different from *qualia,* in that they don't seem to have the qualitative character of the latter and are even more "ethereal" and abstract. Perhaps we should think of them as *yet higher-order* classificatory states.

But now we need to say more about this concept of classification itself, which does so much work in the account I have developed. For it might be objected that insofar as that account rests on the notion of classification, it rests on a notion which is *itself* inherently mentalistic, so that our problem, the qualia problem – or at least the broader problem of the mind's place in the natural world – has not after all been solved at all, but merely kicked up to a higher level: if we explain qualia in terms of classification, and classification is itself a mental process, what explains *it*? To be sure, as I have already indicated, the notion of classification I've been appealing to thus far is not supposed to be understood in mentalistic terms: we are not to think of the classification we've spoken of as involving anything like an *intentional* or *purposeful* putting of external or internal objects and events into classes or categories, or of taking them as having *significance* or *meaning*. Rather, we can understand the "classification" involved in perceptual experience – at least as *so far* spelled out – on the model of the (to all appearances) meaningless processes taking place in a thermostat or even a sophisticated computer as it goes into states which correlate in a law-like way with certain features of its external environment and thus allows it causally to interact with that environment in a way that mirrors (if not, so far as anything said so far shows, literally *duplicates*) intelligent human behavior. The qualia with which we are directly aware and which we have identified with such (classificatory) states are thus so far to be understood as correlated with, but not necessarily literally *representing*, external features, objects, and events. That is,

everything said so far about the subjective, first-person realm of qualia applies to that realm even if understood as a meaningless, uninterpreted cacophony of sensations – what William James called a "blooming, buzzing confusion": think of the conscious experiences of someone whose capacity for higher cognitive processes has been destroyed, so that he is incapable of understanding or attaching any significance to the experiences flooding his mind, or think even of the experiences of some of the lower animals. Consciousness *per se* can arguably come apart from *meaning* – hence the tendency to describe what is peculiar to experiences *qua* experiences as "*raw* feeling" (Kirk, 1994). The problem of qualia is, after all, a different problem from the problem of intentionality, and all I've claimed to have shed light on so far is the former problem.

Still, we do in fact *take* qualia to have significance. Our awareness of them only counts as perception because we *interpret* what we see as a chair, or table, or whatever; and even introspecting qualia and seeing them *as qualia*, apart from their significance as representative of external objects, is a matter of interpretation, of ascribing meaning. It's not just that the brain, in a mechanical sense, "classifies" external objects and internal states; *we*, as thinking things, classify them in the ordinary, *intentional* sense. So even when the *qualia* problem is solved, there is much left to do; indeed, perhaps the *central* problem of mind has yet to be touched. This was certainly the view of Franz Brentano (1995), who famously took intentionality to be "the mark of the mental." Even if this is so, we should by no means conclude that

the *qualia* problem is trivial: centuries of effort have been exerted in trying to solve even it, and as what we've seen shows, a genuine solution might – and if I'm right *does* – require a radical revision of common sense. So solving the qualia problem is no mean feat; not for nothing has it come to be called "the hard problem." Nevertheless, it is plausible to take intentionality as the more fundamental phenomenon in that we must presuppose *it* in trying to account for qualia, and in that the qualia problem wouldn't be a problem for even a system having qualia unless it also had the capacity for intentionality, for *taking* or *construing* qualia *as qualia*, and as requiring an explanation.

Lockwood, in the course of some interesting reflections on the subject of intentionality, goes so far as to suggest that "the problem of taking or construing as is not really a distinct problem from that of consciousness itself" (1989, p. 311). The problems of consciousness and intentionality, in his view, ultimately boil down to the *same* problem:

> I would claim that there is no consciousness, no sentience, without taking as. Some philosophers appear to think that the possession of concepts, as opposed merely to behavioral discrimination, is something that goes way beyond mere sentience, and is perhaps restricted to higher mammals. To me, however, it seems, on the contrary, that it is simply incoherent to suppose that there could be a creature that was aware of certain phenomenal qualities or qualia, without being aware of them *as* anything at all. One source of resistance to the idea

> that even the most primitive sentience must carry with it some minimal conceptualization, is the fact that any concept we humans possess, or at any rate for which we have a word, is likely to be far too sophisticated plausibly to be attributable to the most lowly possessors of consciousness. But of course, this problem arises even at the level of dogs; one can say that the dog is wagging his tail because he construes the ring of the door bell as betokening the arrival of his master. Presumably, however, the dog does not possess *our* concept of 'master' or 'doorbell', and probably not 'arrival' either. Even so, dogs surely have beliefs, and beliefs call for some concepts or other. (1989, pp. 311-312)

Now in taking the problems of consciousness and intentionality to be the same problem, Lockwood does not, I think, mean to imply that consciousness and intentionality are *exactly the same feature*. For they clearly are not: we can certainly separate them *conceptually*, in thought, even if they always appear together in reality, and even if they do so *necessarily* (just as being colored and being extended are distinct features even if they can only be instantiated together).[2] Nevertheless, the link he suggests exists between them does seem quite plausible, and what he says dovetails

[2] This goes contrary to the view of writers like Dennett (1991), Dretske (1995), and Tye (1995) who hold that the problem of consciousness really reduces to the problem of intentionality, in that conscious states are (they allege) really just a sub-class of intentional states, so that an explanation of the latter is *ipso facto* an explanation of the former. There may be an element of truth in this sort of view, though, in that intentionality is the more basic mental phenomenon in the sense that it is presupposed in every explanation of any other mental phenomenon – and in any explanation of *anything*, for that matter. As we'll see, this may imply that an ultimate explanation of intentionality is in principle impossible.

quite nicely with Hayek's account of perception: for the having of a quale is, on that account, the classification of something as related in some way to some other object of experience, that is, it is the placing of something in a *class*; and this is precisely what "taking as" involves. There thus seems to be a natural bridge between consciousness, as construed by Hayekian functionalism, and intentionality. Any classificatory state will not only be a conscious state, but also, at the same time, an *intentional* state. The question before us then, is this: even if the phenomenal or qualitative aspects of such states have been accounted for on HF, what accounts for their intentionality?

Now as all the talk about "*the* hard problem" implies, many theorists today would apparently take it that solving the problem of intentionality is little more than a mop-up job after the qualia problem has been solved. But if so, they would be mistaken. The Hayekian view does in fact have something to say about intentionality as well as qualia – but what it has to say indicates that unlike the qualia problem, the problem of intentionality can *in principle* never fully be solved. McGinn's mysterianism turns out not to be completely wide of the mark.

It might be thought that intentionality is bound to seem inscrutable on the Hayekian view because that view is a kind of functionalism, and thus shares certain relevant deficiencies with other kinds of functionalism. In particular, it might be objected that Searle's (1980) famous "Chinese Room" argument shows that *no* functionalist account of intentionality is possible. But in my view, Searle's argument

fails. Seeing why it fails will provide an opening to understanding the real reason functionalism – and any other view – cannot completely explain intentionality.

Searle asks us to imagine that he, who speaks not a word of Chinese, is locked in a room, the door to which has a slot through which various questions, written in Chinese, are slipped to him on pieces of paper. He has in the room with him a set of Chinese symbols, on little tiles, say, and a rule-book in English telling him what sorts of combinations of symbols to give out in response to the questions put to him. Suppose he eventually so masters the rule-book and the symbols that his responses to the questions, slipped back through the slot to the questioner, are indistinguishable from those of a native Chinese speaker, and that anyone who spoke Chinese and was unaware of what was really going on would assume he really could speak Chinese. In effect, Searle would thereby pass the Turing test for the understanding of Chinese, where the idea of the "Turing test" (named for mathematician Alan Turing) is that any system, even a machine, which produced linguistic behavior indistinguishable from that of a normal human being could be said literally to be intelligent. But in fact (and this is the point) Searle *wouldn't* understand Chinese; his manipulation of the symbols according to their shape or *syntax*, however skillful, wouldn't amount to a grasp of their *semantics* or meaning. He does exactly what a program does, namely manipulate symbols according to syntactical rules – exactly what the computer model of the mind ("Strong Artificial Intelligence" or "Strong AI," as Searle calls it) says the mind does in understanding language – and yet he has no understanding at all. So that model of

the mind must be false. More generally, the functionalist approach to the mind must be false: for the Chinese Room scenario shows (or could be made to show by an appropriate alteration of the details of the thought experiment) that a system can have just the functional organization the *brain* is said to have in going through the "program" that gives rise to linguistic behavior, and yet lack altogether any genuine understanding or intentionality.

The most widely discussed reply to this argument – one discussed by Searle himself in the very article in which he originally presented the argument (in print) – is the "systems reply," which suggests that whether *Searle* understands Chinese or not is beside the point. In the thought experiment, Searle is only one part of a larger system which also includes the rule-book, the tiles, the room, and so forth. He is, as it were, the central processing unit of the system. But just as it wouldn't do to speak of the central processing unit of a computer as running a program, since it is the computer system as a whole which does so, neither is it appropriate to speak of Searle as "running the program" for understanding Chinese, or of understanding or failing to understand Chinese *himself*. It is the *system as a whole*, Searle plus the room and its contents, which is relevantly said either to understand or not understand Chinese; and nothing in what Searle has said shows that it doesn't. Searle's reply to this is to say that the room is irrelevant to the gist of the argument. We need only imagine instead that Searle *memorizes* the symbols, perhaps in their verbal forms this time, as well as the rules governing their combination in response to questions posed to him. In this

case, we can imagine that Searle does exactly what he would do in the Chinese Room, i.e. manipulate symbols on the basis of their shape (or perhaps sound, in this case), and that even if he does so in a way which makes his linguistic behavior indistinguishable from that of a native Chinese speaker, he cannot be said really to understand Chinese, to know the meanings of the symbols.

The first thing I think should be said in response to all of this is that Searle appears to be running together issues which need to be kept distinct. We earlier distinguished the problem of consciousness not only from that of intentionality or meaning, but also from that of *rationality*. The latter problem is that of explaining how material systems are capable of moving from one state to another (in accordance with *causal* laws) in a way that parallels the mind's movement from one thought to another (in accordance with the laws of *logic*). As Fodor has suggested (see the interview in Baumgartner and Payr, 1995), the computer model of the mind, and perhaps functionalism more generally, can be regarded as by themselves explanations only of the *latter* phenomenon. In any case, they can plausibly be regarded as theories of how the brain instantiates states which are related to one another in precisely the way mental states are, but not (without supplementation) as theories of how those mental states get intentional or semantic content. So even to show that the Chinese Room – or Searle himself, if he memorizes the rules, etc. – failed to understand Chinese would be irrelevant to showing the truth or falsity of functionalism. For all

Searle has said, mental states might be functional states, even states of a program; it is a *further* question how those states get their intentional content.

The most popular sort of reply to that further question, one which Fodor himself has presented influential variations on (e.g. Fodor 1987), is that intentionality can be explained in terms of some sort of *causal relation* between functional states and that which they are said to represent. Such a causal relation might be one which gave rise to a law-like connection between an internal functional state and a feature of the external world: a functional state which was instantiated in the brain when and only when a certain external feature caused it to occur could be said to *represent* that feature. Applied to the case at issue, any system that both had the right functional organization *and* the right causal relations to the world *would* have genuine understanding of Chinese. Now the precise details of this sort of story are extremely controversial and extremely complicated, but even in this sketchy form, the idea illuminates what I trust many readers will have sensed already as a significant difference between the two sorts of case Searle describes, namely the Chinese Room case and the memorization case. While it is easy to see how Searle's behavior in the room could be accompanied without any understanding, it is much less obvious that understanding would be absent in the memorization case, at least when you think about it in any detail. To pull the trick off convincingly, Searle would have to memorize so much, and be able to go through it all in his head, without the benefit of tangible aids to memory, with such rapidity, that it is hardly *outlandish* to suppose that

somehow this would have to yield genuine understanding. (Indeed, when you think about the enormity of the task even in the strict Chinese Room case, it is at least slightly less obvious that Searle would lack understanding.) Part of the sense that genuine understanding might fall out of this process after all is, of course, because of the staggering complexity of the mental task itself; but part of it is also because imagining it involves imagining *causal interactions* of a sort which obtain in the case of normal speakers and don't obtain in the strict Chinese Room case. Here Searle wouldn't be responding merely to slips of paper, but rather, in addition to the sounds he hears and symbols he sees, to such factors as vocal inflection, facial expressions, gestures, the immediate environment, and so forth. These causal factors would surely influence his own "processing" in such a way that genuine representation and understanding would plausibly result; words which seemed nothing more than meaningless noises come to be tied to objects and contexts, and through them to *meanings*.

Of course, none of this by itself *proves* anything; and Searle could always insist that we limit even the memorization case to something like the room situation. Perhaps he sits with his eyes closed and answers questions put to him, ignoring the sorts of factors mentioned. But the point is that the minute we imagine the memorization case *together with* all the causal factors, the intuitions Searle is counting on aren't at all as strong as they might have been before thinking about the case in detail. At the very least, I think we can conclude that *Searle's* argument hardly *proves*

that functionalism is false, much less that functionalism together with some causal theory of intentionality is false.

The full vindication of such a combination would, however, have to involve a detailed defense of some kind of causal theory. I have no strong convictions about the relative plausibility of the various types of causal theory and their rivals (e.g. "functional role" theories), though I suspect that, while no causal theory currently on offer is without its defects, much of the truth lies with some such story. But in any case, I will not offer a defense of any particular theory here – that would require a work at least as long as the present one. What I *will* try to argue is that *any* theory is going to fall short of a full and satisfying explanation of intentionality. Any deficiencies in the functionalist theory of the mind, Hayekian or otherwise, in this regard, are deficiencies shared by any possible theory, because they derive from inherent limits on the nature of the mind's understanding of itself. And Hayek's position explains why this is so; so that here too, in the case of intentionality, Hayek takes us as far towards a complete understanding as we are likely to be able to go.

Let's begin by going back to the idea that mental states, in both their phenomenal and intentional aspects, are classificatory states. I have argued that this conception enables us to remove at least much of the mystery surrounding the relationship between the mind and the rest of the natural world. But ironically it might also replace the mystery removed with one that *cannot* be removed. For Hayek holds

that the classificatory character of the mind guarantees that it can never *fully* understand itself, for reasons he sums up as follows:

> [A]ny apparatus of classification must possess a structure of a higher degree of complexity than is possessed by the objects which it classifies; and... therefore, the capacity of any explaining agent must be limited to objects with a structure possessing a degree of complexity lower than its own. If this is correct, it means that no explaining agent can ever explain objects of its own kind, or of its own degree of complexity, and therefore, that the human brain can never fully explain its own operations (1952, p. 185).

Tantalizing and suggestive as this passage is, its import is not entirely clear: what *exactly* are the limits Hayek has in mind, and why are they insurmountable *in principle*?

Part of the answer has to do with Hayek's view that the character of the classificatory activity that constitutes perceptual experience is determined by the ("pre-sensory") experiential history of the individual organism and the evolutionary history of the species. This history shapes the parameters of an organism's possible perceptual experience by hardwiring into the brain the discriminatory capacities that are most conducive to the survival of the species. The neural connections determined by this history and themselves determining perceptual experience and the behavioral dispositions we saw that Hayek thinks is tied to it embody, as noted earlier, a sort of a priori knowledge of certain features of the external world:

> A certain part at least of what we know at any moment about the external world is therefore not learnt by sensory experience, but is rather implicit in the means through which we can obtain such experience; it is determined by the order of the apparatus of classification which has been built up by pre-sensory linkages [i.e. neural connections]. What we experience consciously as qualitative attributes of the external events is determined by relations of which we are not consciously aware but which are implicit in these qualitative distinctions, in the sense that they affect all that we do in response to these experiences (1952, p. 167).

But as this passage implies, this knowledge is not explicit, but *tacit*. What it is that we "know" about the world and how to interact with it is not known consciously. As Miller puts it, "Evolution adapted the eye to facts about optics, but nowhere in the eye can one find a representation or explanation of those facts" (1996, p. 60). We don't know precisely what it is that we know. And since the knowledge in question is what determines the character of the mind's classificatory activity, it follows that we don't know all there is to know about that activity.

What Hayek is arguing is that the explicit "knowledge *that*" something is the case which derives from sensory experience rests on implicit "knowledge *how*" to get about the world, a kind of knowledge which can never be made completely explicit (1952, p. 39). Now the distinction between these two sorts of knowledge and the notion that the former rests ultimately upon the latter is a theme explored in great

detail (though not in precisely these terms) by the later Wittgenstein, especially in his posthumous On Certainty (1969), and has also been dealt with by such prominent thinkers as Martin Heidegger, Gilbert Ryle, and Michael Polanyi. Like those writers, Hayek thinks that this tacit "knowledge how" underlies also our abilities to take what we are aware of in perception as having a certain significance or meaning and to draw conclusions from it; that is, it forms the basis of intentionality and reason (see especially Hayek 1967, and also 1988, Chapter 1). And like them again, Hayek holds that the character of this knowledge is partly determined by *cultural* factors, as well as by biological ones.

The idea can, I think, be made clearer by thinking of it in terms of the problem of rule-following made famous by Wittgenstein. The rules that govern the use of language and logical and mathematical practice, Wittgenstein holds, are determined by "forms of life" or sets of cultural practices that communities simply take as given, as what determine what is legitimate and illegitimate but are not themselves subject to evaluation as to their legitimacy (Wittgenstein 1953). Now whether the relevant "community" is supposed to be a given local human culture or the human race as a whole will to some degree determine whether and to what extent all this is given a relativistic-cum-skeptical reading, as will the answer one gives to the question of why exactly some practices and not others are taken as given. What Wittgenstein's own view of these matters was is, of course, a subject of great controversy. Hayek's view, however, is clear. The practices in question, which embody the rules that govern

language and reason, are determined by *cultural evolution* as much as by biological evolution.

Hayek's notion of cultural evolution is one according to which those practices which best enable a group of human beings to adapt to its environment will be those which survive, for the groups that practice them will be those which proliferate and keep these practices alive; while those practices which are ill-suited to the preservation of a group will die out, since either the group that practices them will itself shrink or die out, or will abandon those practices and adopt those of more successful groups. That the practices in question will in fact facilitate the adaptation of a group to its environment is not necessarily the reason why the practice is chosen; indeed, it rarely is, for that the practice has this utility is usually only discoverable after the fact if at all. The practice may in fact be chosen for reasons that have no relationship to its actual value, perhaps even for superstitious reasons. But this is irrelevant to the causes of the practice's preservation, as well as to its actual value. (Compare the situation in biological evolution, where a feature comes about, not because it is advantageous to the organisms possessing it, but because of a random mutation; rather, it is because it *turns out* to be advantageous that it is preserved or selected for and is of value to the organism.)

Applied to the rules that form the basis of intentionality and reason, Hayek's claim is that those rules that aren't hard-wired into the brain as a result of biological evolution are inculcated by means of this process of cultural evolution. That many of

these rules appear to be widespread, even if this can be argued to be not the result of biological evolution, can be accounted for by Hayek's notion of cultural evolution, since it is likely that groups which follow non-adaptive rules will either shrink or die out or abandon those rules. Fodor has argued that "Darwinian selection guarantees that organisms either know the elements of logic or become posthumous" (1981, p. 121); Hayek might add: "And even if it doesn't, cultural evolution will." This account of things preserves, I think, the insights of Wittgenstein's account while avoiding the latter's potential weaknesses. For Wittgenstein's "forms of life" arguably amount to arbitrarily chosen practices that have no necessary connection to the way the world is: relativism and skepticism loom. Not that Wittgenstein himself intended such a result; I am quite sure he did not, but his account is ambiguous enough – or, more charitably, subtle enough – that it is very easy to take it in this direction, and scores of pseudo-Wittgensteinians of the "sociology of knowledge" stripe have done just that. In any case, on Hayek's account, that certain practices are preserved is the result of the adaptive advantage they provide (regardless of whether those practicing them are aware of this); and that they provide this advantage is certainly strong evidence for their corresponding to some extent to the way the world is.

But again, the thing to keep in mind here is that these rules are not necessarily consciously chosen for their utility; and in fact their utility might be quite unknown or even unknowable. Indeed, they are not necessarily consciously *chosen* at all. We just abide by them, without realizing it. The rules by which we perceive, understand, and

reason about the world are not themselves perceived or understood by us, nor did we arrive at them by a process of reasoning. They are inculcated by biological and cultural evolution. As Hayek sums it up: "Mind is not a guide but a product of cultural evolution, and is based more on imitation than on insight or reason" (1988, p. 21); and "It is less accurate to suppose that thinking man creates and controls his cultural evolution than it is to say that culture, and evolution, created his reason" (p. 22).

Because the rules which govern its operations – which govern perceiving and reasoning – are not consciously chosen or known but are presupposed in all conscious activity and all knowing, the mind does not fully understand itself. But even if this is in fact true, need it be? Couldn't we come to discover these rules and state them explicitly, thereby attaining a full understanding of ourselves? Hayek answers that this is impossible. For even if we come to understand some of the tacit knowledge that guides our mental processes, this understanding itself would be governed by yet higher-order rules which would remain tacit or inexplicit:

> It is important not to confuse the contention that any such system [as the mind] must always act on some rules which it cannot communicate with the contention that there are particular rules which no such system could ever state. All the former contention means is that there will always be some rules governing a mind which that mind in its then prevailing state cannot communicate, and that, if it ever were to acquire the capacity of

> communicating these rules, this would presuppose that it had acquired further higher rules which make the communication of the former possible but which themselves will still be incommunicable (1967, p. 62).

Hayek's claim here can, I think, be illuminated by comparison with the notion of "the Background" (of tacit knowledge) developed by John Searle (1983, Chapter 5; 1992, Chapter 8). Searle argues that intentional mental states – beliefs, desires, and the like – have the content they do only by virtue of their place in a vast network of intentional states: the desire to run for the presidency of the United States, for example, has the intentional content it has only in the context of other intentional states such as the belief that the United States has periodic elections, the desire that voters cast their votes for one, and so forth; and if the other intentional states were different, the intentional content of the desire would be different. But this network itself functions against a background of capacities which are themselves non-intentional, non-representational. The sort of "capacities" Searle has in mind are essentially the things we have been calling pieces of tacit knowledge, i.e. the presuppositions of everyday conscious and explicit reasoning which are rarely or never themselves made explicit or consciously considered. And because they aren't, they aren't, strictly speaking, intentional or representational at all. Commonsense realism about the external world is, Searle says, an example of such a capacity, something that isn't really a belief but a presupposition of our beliefs:

> My commitment to 'realism' is exhibited by the fact that I live the way I do, I drive my car, drink my beer, write my articles, give my lectures, and ski my mountains. Now in addition to all of these activities, each a manifestation of my Intentionality, there isn't a further 'hypothesis' that the real world exists (1983, pp. 158-9).

In other words: all thought and action proceeds *as if* it consciously presupposed an explicit belief in the reality of a world outside the mind. But in fact such a belief isn't explicit, and not even *implicit*, strictly speaking; for (at least generally) it isn't really a "belief" at all. I don't ***believe*** that the external world exists: I simply *act* in a way that makes sense only given that there is one.

Now that the Network of intentional states rests on a Background of tacit "knowledge" (Searle capitalizes the terms to signify their status as technical terms) is true not only *in fact*, but of *necessity*, in Searle's view. For since the intentional states which make up the Network get their content from other such states, if there were no non-intentional Background, then in tracing the links that give any particular intentional state its content, we would be led into an infinite regress (1983, pp. 152-3). Even if, in trying to undertake some activity, I consciously follow explicitly formulated rules, those rules themselves are capable of various interpretations; and the same is true of any further rules I might appeal to in order to interpret the first set. So ultimately, I *must* simply *act* in accordance with some interpretation of some set of rules, without explicitly or consciously choosing to do so; otherwise I would never get

started.[3] (And though Searle doesn't give this example, we might also think of Lewis Carroll's famous parable "What the Tortoise Said to Achilles" (1977, pp. 431-4), in which the hapless Achilles finds that he is unable to proceed in running through a simple modus ponens argument if he tries explicitly to formulate each assumption lying behind the inference.)

But if all understanding thus *must* rest ultimately on *action* – on simply taking things to have a certain significance – then it is clear that intentionality can never fully be explained: a *full* explanation would have to be an explanation of this ultimate act of "taking as" itself, and that, in the nature of the case, is impossible. Of course, it might be explained in *one* sense. We can perhaps find some explanation for why such and such a *particular* level in fact serves as the stopping point, maybe in terms of either cultural- or even biological-evolutionary factors which hard-wire certain assumptions into us. But their character *as assumptions*, as having a particular intentional content or significance, cannot be explained. *Why* do I take things to have a certain significance? At the end of the day, the only answer possible is: *I just do*.

It should thus be no surprise if the various suggested accounts of intentional content, causal and otherwise, seem inadequate. No doubt some kind of causal connection between a mental state and what it represents is a *necessary* condition for its representing that thing, as writers like Putnam (1975, 1981) have plausibly argued.

---

[3] This is by no means the only sort of argument Searle gives for the hypothesis of the Background, but it is the one most similar to the sorts of considerations Hayek has in mind. For Searle's full defense of this hypothesis, see his 1983, pp. 144-153 and 1992, pp. 178-186.

But it seems hardly a *sufficient* condition. Suppose the reason my current thought that there's a computer screen before me has the intentional content it does is because of its causal connection to the screen itself, etc. Knowing this might tell me why it has that *particular* content as opposed to the content that there's a cat in front of me; but does it tell me why it has *any* content *at all*, why I'm able to take my thought *as* a thought of this or that sort? It seems not. As Searle writes of causal theories, the problem is that it appears possible that a system could have all the causal relations such theories speak of, and yet lack intentionality (1992, p. 51). He takes this to indicate that such theories have no value. That seems to me to be too strong a conclusion – we need suppose them at worst incomplete. But I think he is on to something when he goes on to say that the problem with such accounts considered as complete explanations of intentionality is that they try to *reduce* intentionality, which is an inherently *normative* notion, to non-normative, "brute" elements. That is, they try to reduce the meaningful to the meaningless, and in effect leave out the meaningful altogether.

Of course, this parallels the failure of reductive accounts of *qualia*: intentionality, like qualia, is simply irreducible to physical processes understood in the standard, naïve way. Now there is an extent to which the strategy of appealing to indirect realism and structuralism helps even here: if the worry is that the physical world seems intrinsically devoid of intentionality, so that there is no room within it to fit the mind, the response is that we lack any knowledge of the intrinsic nature of the

physical world which would justify such a worry, and that what we do know most intimately is in any case precisely that portion of the physical world where we know intentionality exists, namely the mind/brain. The problem of intentionality is thus *not* a problem about how intentional states can be identical with states of the brain.[4] But removing doubts about such an identity hardly explains what *gives* an intentional state its intentionality.

These remarks also suggest a respect in which even the qualia problem is not *completely* solvable, at least insofar as that problem goes beyond merely explaining how qualia can be identical with states of the brain. We can explain why any particular quale has the character it does in terms of its relations to other qualia, just as we can explain the intentional content of a given mental state in terms of its relations to other mental states. But just as in the case of intentionality we must ultimately come to a point where we just take meaning or intentional content as such as given, so too in the case of qualia, we must take the fact of *qualitative* content as such as given. I can say why this or that quale has the character it has in terms of its relations: I cannot say anything very interesting about why it has *any* such character at all.

Incidentally, something similar seems to go for rationality, at least insofar as this involves an appeal to rules of inference, standards of justification, and the like which themselves seem in need of justification. Of course, the evident *necessity* – the

---

[4] The difficulty posed by intentionality thus hardly supports dualism. It is a difficulty faced by *any* attempt to understand the mind, and has nothing to do with the question of what sorts of substances or properties there are in the world. The problem is equally present whether intentional states are

impossibility of the contrary – of the laws of logic and of mathematical truths plausibly justifies *them*. (Though intuition-manipulating Cartesian evil demon scenarios might play merry hell even with our confidence about this!) But in the case of induction, say, or the reliability of memory, or perhaps even the trustworthiness of the senses (if the sort of defense of belief in the external world considered earlier is found wanting), we seem to be at a loss when asked for a rational justification. Again, we can of course explain why we take these to be fundamental to rationality in an *evolutionary* sense, since given what we know (or take ourselves to know) about the world, it seems highly unlikely that creatures who lacked at least a tacit adherence to these principles could survive; but this can't serve as a *rational justification* of such principles – not a non-circular one, anyway, since we have to assume them in order to get an evolutionary argument, or almost *any* argument at all, going. (All this no doubt accounts for the fact that we both cannot seem convincingly to refute skepticism, but also cannot seem to take it seriously either.) Any attempt to explain intentionality must ultimately *presuppose* intentionality; any attempt to explain rationality must ultimately *presuppose* rationality.

In a way, none of this should be surprising if we take seriously the picture of our knowledge of the world shared by both the Russellian and Hayekian views. Our epistemic *starting point* is, in every way, our knowledge of the realm of the mind, which is the realm, not just of consciousness, but also of intentionality and reason. We

---

construed as immaterial or physical: in either case, we are left ultimately unable to explain fully *why*

know these better than anything else, so that it is to be expected that we have a hard time finding anything more basic in terms of which they can be explained. The Hayekian view reveals, as I have tried to show, that there really is no barrier after all to seeing how the mind can be part of the natural world, since we only know the mind itself with any clarity and know the rest of the world only *through* it; but it leaves the mind itself, in all its facets, rather *sui generis*.

In any case, the way this all fits in with, and is reinforced by, Hayek's general functionalist account of the mind – and with the rather cryptic passage about the limits of the mind's self-understanding cited earlier – is as follows. As we've seen, perceptual experience is, on Hayek's account, just the brain's classificatory or differentiating activity in response to the stimuli impinging upon it. This activity consists in the forming and strengthening or weakening of neural connections and sets of neural connections, different sets of neural connections corresponding to different attributes of a stimulus and perception of the stimulus amounting to the "superimposition" or co-occurrence of impulses in the various connections corresponding to its attributes. Perception of stimuli thus requires that there be a larger number of sets of connections corresponding to various possible attributes than there are stimuli – to perceive even a single object like an orange, for example, I must possess multiple sets of neural connections, corresponding to orange-ness, roundness, and the like. Now this process is constrained by the evolutionary history of the species

---

we are able to attach significance to those states.

and the past history of the individual, the latter partly consisting of the inculcation of cultural practices that give perceptual experiences their cognitive significance. These constraints amount to tacit rules that determine whether a stimulus is to be classified one way or another (i.e. whether perceptual experience is going to have this quality or that), what behavioral responses to stimuli we are disposed to, and what inferences we are disposed to make from our experiences; and if these rules become explicit, it is only because of the operation of higher-order tacit rules.

Such rules consist ultimately in just the existence of certain higher-order neural connections which govern the classificatory connections that constitute experience, though, and the making explicit of them just amounts to the forming of yet higher-order classificatory neural connections. And as in the case of perception of external stimuli, this is a matter of the superimposition of connections corresponding to different aspects of the rules. So again, there must be a larger number of possible sets of connections corresponding to possible attributes of rules than of rules themselves. This idea is what Hayek has in mind when he says, in the passage quoted earlier, that "any apparatus of classification must possess a structure of a higher degree of complexity than is possessed by the objects which it classifies" (1952, p. 185). And it follows from it that it is impossible for all the rules that govern the mind to be made explicit; for to make explicit or classify all the rules governing it, the mind would have to be more complex than itself (1952, Chapter 8, Section 6, passim). The most we can attain is thus an "explanation of the principle" on which the mind operates (1952,

p. 182), an understanding of the fact that it involves the following of rules of the sort described; we can never have an explanation which makes explicit the *details* of the process, a *spelling out* of all of those rules.

To anyone familiar with the recent history of research in artificial intelligence and cognitive science generally, much of all this might have a familiar ring. The problem Hayek says faces any attempt fully to understand the mind is reminiscent of what among AI researchers is called the "common-sense knowledge problem." That problem is one of discovering how a machine can be designed so as to instantiate the sort of common sense assumptions about the world that underlie everyday thought and practice in human beings, assumptions which go too deep for us ever to be conscious of them to any great extent. In order for a machine genuinely to be said to be intelligent, it would not only have to be able to carry out an extended conversation, Turing test style, on some particular occasion or occasions; it would have to reflect in its linguistic behavior the tacit knowledge human beings have and which makes them capable of adapting to unforeseen circumstances. A computer program designed to produce intelligent sounding responses to questions put to it about gardening, and thus conjoined with a database containing encyclopedic information about various types of weeds, fertilizers, and so forth might indeed fool even the greenest of thumbs among human interlocutors – until asked, say, whether a certain weed is known to play the ukulele. Such an odd query would produce puzzlement in a human being, of course; but he'd be able to respond, if convinced the question was serious, "Of course not!

Are you crazy?" He would, in a sense, *know* that there are no ukulele-playing weeds, but not because he'd ever learned it explicitly. Such knowledge is just part of what Searle would call "the Background" of human thought; and this is just one example among potentially *millions* of similar examples. But even a very sophisticated gardening-discussion program wouldn't know what to do with such an unlikely question, unless somehow prepared in advance for it, the fact that weeds don't play ukuleles having been fed into its database. For a machine to instantiate genuine intelligence, then, it would have to be capable of just the sorts of reactions human beings would give to such examples. But the problem is that this would seem to require that the machine be programmed with millions of such arcane bits of knowledge, over and above the millions of more obvious and interesting pieces of knowledge (e.g. standard facts about gardening) we'd already assumed it would need to have; and it's unlikely that even carrying out such a programming task would prepare it for *all* of the odd situations human beings would be able to deal with without difficulty. This, at any rate, is how things would have to go on the standard, "symbolic processing" model of computation, on which the knowledge built into the system and against the background of which it processes symbols must be explicit. It seems just unlikely in the extreme that all of the common sense knowledge we take for granted can be explicitly represented in the manner required. More importantly, even the knowledge such a system would have would itself be subject to varying *interpretations*. That is, the facts programmed into it would have whatever

significance they have for the system only given certain other facts; and ultimately, for reasons we've seen, it appears that significance must rest ultimately on *in*explicit, *non*-representational dispositions to *act*, not explicit representations of the sort foundational to the symbolic processing model.

This problem has received a great deal of attention in recent AI research, and it is partly due to the arguable inability of the symbolic processing model to deal with it that many researchers have adopted instead the "connectionist" (or "neural networks" or "parallel distributed processing") model of computation. On this approach, the mind is thought of, not as a system which processes explicit symbols in serial fashion and according to fixed syntactic rules, but rather as a dynamic network of connections between sub-symbolic nodes or units having tendencies to excite or inhibit each other, in parallel fashion, and according to degrees of strength (or "weights") between connections which vary as the whole system evolves over time in response to new inputs to the system. (Less abstractly, we might think of the system on which this model is based, namely the brain itself, a system the nodes of which are firings of neurons and groups of neurons, which have tendencies to excite or inhibit one another according to the strengths of the connections between them that have evolved as the organism interacts with its environment.) In a connectionist system, it is not localized, explicit symbols which act as representations, but rather patterns of excitation or inhibition instantiated across the system as a whole, giving the representations existing in such a system a dispersed, inexplicit quality. This inexplicit character of

connectionist representations is thought by many more adequately to model the inexplicit, tacit character of common sense knowledge, and a connectionist system's tendency to *evolve* and revise its representations (as a result of the shift in weights between connections, and so forth) in response to new situations rather than to operate according to fixed rules is likewise thought better to model the human ability to respond effectively to unforeseen situations. And of course, given the evolving character of such a system and the fact that it operates ultimately according to whatever tendencies, instantiated in weights between connections, serve as most basic at any given moment, it also follows that the representational character of the system, the precise *content* of its representations, can only be understood in the most general way and never in explicit detail.

With all of this in mind, it is hardly surprising that Hayek's work should be thought to be a precursor to the connectionist paradigm.[5] That paradigm appears to model the mind in a way that is sensitive to the sorts of limitations on it's self-understanding that Hayek takes to be inevitable, and which the orthodox symbolic processing model can plausibly be accused of ignoring. It also, incidentally, shows that recognition of those limitations need in no way commit one to the conclusion that the mind cannot be understood in computational terms of any sort, much less that it cannot be taken to be a kind of physical system. Hayekian considerations – and the

[5] It also for similar reasons has been compared to the "complex adaptive systems" research associated with the Santa Fe Institute, especially John Holland's (1992a, 1992b, 1995) work on genetic algorithms and what he calls (in another obvious parallel to Hayek) "classifier systems." See Miller (1996).

common sense knowledge problem to which they are related – show at most that the mind cannot be regarded as a *classical, symbolic processing* type computational system, not that it cannot be *any* sort of computational system.[6] Not that it shows even this, at least not without qualification: as a number of writers have speculated, it might be that the right computational model of the mind will turn out to be one that incorporates elements of *both* the symbolic processing and the connectionist approaches.

The position under consideration, given its appeal to the notion of inexplicit, higher-order rules which are presupposed by all explicit understanding – rules which stand, as it were, outside the system – might also bring to mind Kurt Gödel's famous incompleteness results in mathematical logic. Indeed, Hayek himself suggested that "Gödel's theorem is but a special case of a more general principle applying to all conscious and particularly rational processes, namely the principle that among their determinants there must always be some rules which cannot be stated or even be conscious" (1967, p. 62).[7] Nevertheless, it should by now be clear that whatever the

---

[6] Indeed, this seems to be the best interpretation of most of the influential criticisms of AI. Dreyfus and Dreyfus (1990), for instance, famously object to the classical paradigm precisely because of its apparent inability to deal with common sense tacit knowledge, but they are nevertheless at least open-minded about the possibility that the connectionist approach isn't subject to the same objections; and even Searle has nice things to say about connectionism (1992, pp. 246-247). At the same time, it doesn't follow that anyone sympathetic to the general connectionist approach need accept any of the *specific* connectionist models of the mind now current. Indeed, Hayek, like other writers sympathetic with connectionism (e.g. Dreyfus and Dreyfus 1990, p. 330), suggests that even these models may be inadequate; for it may turn out that for a model adequately to represent a complex system such as the brain, it would have to amount to, not merely a *model*, but a complete *reproduction* of that system (1982, pp. 292-293).

[7] Hayek also suggested that his claim that any mechanism of classification would have to possess a greater degree of complexity than what it classifies "would seem to follow from what I understand to

connection between Hayek's position and Gödel's work, the former is, again, *not* intended to show that the mind cannot be understood in computational terms or that it cannot be a material system. It is not to be confused with the controversial – and in my view, as in that of most theorists, implausible – suggestion, advanced originally by J.R. Lucas (1961) and recently revived by Roger Penrose (1989) that this is precisely the lesson to be derived from Gödel's results.[8] It is arguable, though, that Gödel's results cast doubt on the adequacy of a *purely symbolic processing* approach to the understanding of the mind and toward the attempt to reproduce intelligence in machines. As Rudy Rucker notes, Gödel's work shows that:

> [T]he human mind is incapable of mechanizing all of its mathematical intuitions. For to mechanize our intuitions is to produce a finite description of a formal system *K*. But as soon as we see this finite description, our mathematical intuition shows us a fact, Con(*K*) [the statement that *K* is consistent], which the mechanized system does not prove. So it is not true that the mechanized system *K* proves all facts that we can perceive through our mathematical intuition. (1982, p. 180)

---

be Georg Cantor's theorem in the theory of sets according to which in any system of classification there are always more classes than things to be classified, which presumably implies that no system of classes can contain itself" (1967, p. 61, n. 49). We might also note that the idea that there is in principle no way completely to enumerate the rules according to which the mind operates is reminiscent of what, according to Patrick Grim, follows from Cantor's results, namely that "there is no set of all truths" (Grim 1984). But this is all a bit vague. Fully to spell out the connections between Hayek's work and that of Cantor (and that of Gödel for that matter) would require a study of its own; and this is not that study.

[8] See Searle 1997, Chapter 4 for one response to this suggestion.

But if human-like mathematical intuition is not mechanizable, then we could never write a program of explicit rules by virtue of the following of which a machine would duplicate exactly our powers of mathematical reasoning. Nevertheless, Rucker says, "even though we cannot write the program for a theorem-producing machine that is equivalent to human mathematical intuition, it is possible that such a machine could exist and even be empirically discoverable" (1982, p. 180). How? "The answer is evolution" (p. 181). Specifically, Rucker describes a process inspired by John von Neumann's theory of self-reproducing automata, whereby robots are constructed to run relatively primitive programs directing them to create copies of themselves, but which programs are also designed to produce occasional *mutations* in succeeding generations. Natural selection then operates on such mutations in the basic program (along with, say, alterations due to shuffling of sub-programs between robots if we also factor in some kind of sexual-reproductive aspect to the process), until, after many generations, cognitive capabilities – including mathematical ones mirroring our own – are produced the computational underpinnings of which would then be too complex for us to model and which could thus not possibly have been put into machines by fiat. But however, spelled out, whether in this fashion or along the lines of the evolution of a connectionist network, the Gödelian notion that the mind, even if instantiated in a physical system, cannot simply be created by fiat according to explicit

rules but must *evolve* according to principles which forever remain largely inexplicit, parallels exactly Hayek's own results (and I think helps to elucidate those results).[9]

These, then, are the sorts of considerations that lead Hayek to conclude that the operations of the mind, particularly in respect of their intentional and cognitive aspects, rest on a foundation of tacit knowledge which cannot, in principle, be made fully explicit. And since it cannot, it will forever be impossible for us completely to understand those operations. We are simply unable to get outside our own skins, as it were, and survey the systems that constitute our minds; for we *are* those systems. This dovetails with the indirect realist, structuralist, and Kantian aspects of Hayek's position: our conception of the world – including ourselves – is unavoidably conditioned by built-in constraints, and of necessity, we can't step outside those constraints, see what the world is like independently of them, and note just *how* they

---

[9] Hayek himself was aware of the affinity of his approach with von Neumann's work, and saw in him one of only a small few who were interested in or grasped the problems with which he was dealing: "[W]hen I was writing The Sensory Order, I reasoned that I could explain to people what I was doing. Usually I found it very difficult to make them understand. And then I met John Von Neumann at a party, and to my amazement and delight, he immediately understood what I was doing and said that he was working on the same problem from the same angle" (Weimer and Hayek, 1982, p. 322). The physicist Erwin Schrödinger was another: "To my great surprise, he was the one man who seemed to have fully understood The Sensory Order. But of course he was working on just this sort of problem" (Hayek 1994, p. 139). But though Hayek planned on exploring further the problem of "what we can say 'within a system' and what we can say 'about a system'" (1994, p. 29) and had begun a paper on the topic, "when he found that no one could follow his discussion, he gave it up" (p. 29) and turned again, and for the rest of his career, to the problems of economics and political philosophy which had always been the main focus of his attention. This pattern of understanding and sympathy on the part of a few eminent scientists and neglect on the part of almost everyone else, including his own immediate colleagues, parallels exactly the reception accorded Russell's work on the mind-body problem. As Lockwood notes, "in Russell's own lifetime, his writings were, for example, widely read and admired by scientists, including Einstein. The physicists Sir Arthur Eddington... and Sir James Jeans... read, understood, and agreed with Russell's views on mind and body. Professional philosophers, however, read his exposition of these views, if at all, through the distorting lenses of their own philosophical preconceptions and have mostly made nonsense of them..." (1989, p. 157).

condition our grasp of that world; and even if we could, it would only be by virtue of guidance by further constraints which we would not thereby have stepped outside of.

(b) The inscrutability of matter: from the sensory order to the physical order

I speculated in the last chapter that it might turn out that *all* our knowledge is and can only be knowledge of structure, never of intrinsic qualities. Whether this is so of every domain, it is certainly true of the mental and physical orders if the account defended here is correct.[10] And the *nature* of our knowledge of these domains – embodied as it is in the instantiation of classificatory states – suggests a new way of spelling out the general relationship between mind and matter. It turns out that we can, after all, see them as of fundamentally the same sort, without having to "reduce" one to the other. We have good reason to identify the mental realm with a portion of the larger physical world (*not*, it can never be too often repeated, understood in common sense or physicalist terms), but since we lack knowledge of the intrinsic nature of either, there are no grounds for taking the physical as in any interesting sense metaphysically more fundamental. So we might, after all, think of the Hayekian view as a kind of "neutral monism," indeed a version more plausibly "neutral" than previous versions, since it doesn't claim to be able to identify the neutral "stuff" out of which

[10] Though again, I refer the reader back to the qualification made earlier in light of Galen Strawson's view that we may well plausibly be thought to grasp in part the intrinsic nature of space. And perhaps in general we can be said to know the intrinsic nature of the objects of geometrical, and all mathematical, knowledge – though of course, these objects are the paradigm cases of *abstract* objects. So even if we know the intrinsic nature of the mathematical realm, it's not clear that this is in any interesting way an exception to the general rule that we can know only abstract structure, for abstract structure just seems to be the "intrinsic nature" of mathematical objects! I will say a little more about this sort of issue presently.

mind and matter are composed: sense-data (especially construed as possibly existing unsensed) were the standard candidates for such "stuff" in neutral monism as developed by Mach, Russell, et al., and many commentators have thus taken the resulting position to be less neutral than phenomenalistic or even idealistic.[11] A mental state is just a particular kind of physical state, namely a physical state that comes to be in a higher-order, classificatory relationship to other physical states. But the intrinsic nature of such states, whether physical or physical-cum-mental, is something we have no knowledge of.

For this reason, it turns out there is no *ultimate* intrinsic difference between subjective and objective realms, and thus no ultimate metaphysical cleavage between the realms. Subjective and objective states are of the same intrinsic sort – whatever that might be – and they are subjective insofar as they are in a classificatory relationship to other such states, objective if they are the object of such classification. The subjective/objective distinction is thus really an *epistemological* distinction of sorts, an artifact of the occurrence of classificatory states in the world: if there were no such states, there would be no "subjective" realm; but this doesn't privilege the objective realm, because in that case, there'd be no "objective" realm either. Subjectivity and objectivity aren't intrinsic properties of the world; again, we don't know any such properties.

---

[11] And given that neutral monism was typically spelled out in terms of sense-data understood as absolute, intrinsic qualities, Hayek himself disavowed the label, 1952, p. 176. As Gray notes, however (1998, p. 166, n. 10), neutral monism is compatible with Hayek's position if not fleshed out in those terms.

Might it turn out that there *are* no such properties – that *all there are* are relations, that the universe *just is* an abstract structure with nothing "fleshing it out"? Bizarre as such a view seems, there are those who defend it or something like it (e.g. the physicist John Wheeler's "it from bit" conception of the physical world as consisting ultimately of information states). Chalmers calls it the "pure causal flux" view (1996, pp. 153, 302-304), and objects that we know there *are* in fact more than just relations, that there are intrinsic qualities – namely qualia, which are what he, in Russellian fashion, suggests might be what fleshes out the causal structure of the universe. But if the view defended here is correct, qualia *aren't* intrinsic properties, in which case this objection to the "pure causal flux" view fails. And that view might arguably have certain advantages other views lack: John Barrow (1992, pp. 280-284) has suggested that it might help explain our capacity for mathematical knowledge. The standard objection to mathematical Platonism, the view that mathematical objects are in some sense real, though abstract, entities, is that knowledge appears to require a causal connection between the knower and what is known, and such a connection appears impossible between the mind and abstract objects like numbers. But if both the mind and the physical world are *themselves* abstract structures, the difficulty would seem to be removed: for we already know that *they* can interact causally, and if the mathematical realm is just another abstract realm, there seems no reason to doubt that the mind could interact with it too.[12]

[12] In its vision of a purely abstract world, such a view would also have the advantage of making it at

Of course, such a radically eccentric view would need careful fleshing out (if you will) before we could regard it as more than merely an interesting speculation. And there is another, more important difficulty facing it (as Chalmers points out), namely that it is arguably *incoherent*; surely, we want to say, relations have to be relations *between* things, and no causal structure could actually exist unless it were the causal structure *of* something existing over and above it. To be sure, Hayek himself took it for granted that the existence of a causal structure presupposed elements which flesh it out (1952, p. 47).

In any case, the Hayekian view does seem to imply a much broader set of possibilities in the domain of general metaphysics than would otherwise be evident. Given the limitations on our knowledge entailed by that view, the dogmatic pronouncements about what ultimate reality must be like (or at any rate, what it must *not* be like) made by positivists, materialists, and others have much of the wind taken out of their sails. And given especially the stress on the *abstract* character of our knowledge, Platonism in its various guises seems much more palatable, whether or not Barrow's suggestion is taken seriously. (And – who knows? – perhaps even theism and realism about values are helped at least slightly by such considerations, though Hayek, probably rightly in my view, accepted neither.)

At the same time, those limitations, given their largely *necessary* character, while undermining dogmatism, also tend to dash hopes for a complete and transparent

---

least *somewhat* clearer what the Pythagoreans could possibly have meant by saying that reality was

understanding of the world. And largely because of results concerning complex systems (like Gödel's results and those of the connectionists) mirroring Hayek's own, this conclusion seems to be gaining currency anyway. Even Dennett appears representative of this point of view, as indicated by an exchange reported by John Horgan:

> "There's a curious paradox looming" in modern science, [Dennett] said. "One of the very trends that makes science proceed so rapidly these days is a trend that leads science away from human understanding. When you switch from trying to model things with elegant equations to doing massive computer simulations... you may end up with a model that exquisitely models nature, the phenomena you're interested in, but you don't understand the model. That is, you don't understand it the way you understood models in the old days." A computer program that accurately modeled the human brain, Dennett noted, might be as inscrutable as the brain itself... He thought a theory of the mind, although it might be highly effective and have great predictive power, was unlikely to be intelligible to mere humans. The only hope humans have of comprehending their own complexity may be to cease being human... [by becoming] able to abandon our mortal, fleshy selves and become machines... but Dennett seemed to doubt whether even superintelligent machines would ever fully comprehend themselves. Trying to know themselves, the machines

---

ultimately composed of numbers!

> would have to become still more complicated; they would thus be caught in a spiral of ever-increasing complexity, chasing their own tails for all eternity. (Horgan 1996, pp. 179-180)

Of course, by itself, the currency of such ideas proves nothing: science, at least as packaged by journalists, is as subject to airy trend-mongering as any other field (and as any of its readers knows, Horgan's book is airy trend-mongering *par excellence*). I note it only to underline the fact that much of Hayek's position, already independently defensible, finds further support in work that has appeared, in a number of fields, since the time he wrote. One needn't approach the issues we've been examining in this chapter from the point of view of the qualia problem (or even think, as McGinn does, that *that* problem is unsolvable) to suspect that there are surprising limits on what we can know.

(c) The inscrutability of man: from philosophy of mind to ethics, economics, and political philosophy

No study of Hayek's philosophy of mind can end without saying something about the important relationship it has been claimed to bear to his better known work in the social sciences. It might seem strange to suggest that there plausibly *could* be such a relationship – surely there can be no interesting connection between one's preferred solution to the mind-body problem and, say, his take on the state of public schools, or the death penalty, or the minimum wage! And yet Hayek's famous defense of the free market and limited government is often said by commentators on his work

to rest on his work in philosophical psychology. John Gray goes so far as to claim that "it is Hayek's view that the impossible ambitions spawned by contemporary culture arise from a false understanding of the human mind itself" (1993, p. 33) and that "socialism and interventionism... are but long shadows cast by a false philosophy of mind" (p. 36)! So are we to believe that lurking inside every dualist or physicalist is a radical waiting to get out?

Of course, that's not quite what Hayek or his commentators have in mind. But what, then? It seems to me that Hayek himself makes less of the alleged connection between the two aspects of his work than some commentators do, but there are indeed some interesting points of contact.

We might first consider the fact that on Hayek's view, the mind is a complex system governed by principles we are incapable of understanding in their entirety, and which consequently is not the sort of system that can be reproduced by fiat. It is also a system which is best thought of as *decentralized*, a connectionist machine whose various subsystems operate in parallel rather than in serial fashion, and without centralized units of significance but rather "distributed representations" (to use a bit of connectionist jargon). Were we to try to reproduce minds like our own in machines, we could thus not do so directly but would rather have to *evolve* them, as it were, in such a way that the results could not be known or planned in any detail.

In these respects, the mind is very much like other complex systems, such as other biological phenomena, the weather, and – as is increasingly understood, in no

small part due to Hayek's own work – *economic* systems. The general structure of the mind is very similar to that of a market economy, and as the work in economics of Hayek and his teacher Mises has shown, the superiority of the market economy over a centrally planned economy is due precisely to its decentralized character; in fact, as Smith (1996) notes, the differences between the two parallel very closely the differences between the connectionist model of the mind and the symbolic processing paradigm. That parallel is most clearly seen in Hayek's (1997) account of the role prices serve in transmitting economic information.

Hayek elaborates on the function the price system serves by calling our attention to the dispersal of knowledge in a complex society like our own. The information relevant to the determination of the most efficient allocation of resources isn't located in any central location or accessible to any single mind. Rather, it is dispersed among millions of individuals, each of whom is intimately familiar with the circumstances of his own time and place, but largely ignorant of the circumstances prevailing in other parts of the economic system. Moreover, this information *cannot* be centralized, cannot be put together for the perusal of, say, a socialist central planning board. For not only is it fragmented and dispersed in such a way that gathering it together is a practical impossibility, but much of it is *fleeting*, that is, it is information about local circumstances that rapidly change, so that even if such information could be gathered, it would largely be obsolete by the time the gathering process was completed. And in addition, much of the relevant knowledge isn't

propositional knowledge at all, isn't knowledge of data that could be recorded in a ledger or fed into a computer; but is rather what we've been calling "tacit" knowledge, in this case knowledge which is embodied in the habits, practices and conventions of business life. (Think of the distinction between "book-learning" and the "know how" which derives only from hands-on experience, where the latter, by its very nature, cannot be communicated in an explicit way.)

But though it cannot be centralized, as socialism would require it to be, this information nevertheless gets utilized in a competitive market in such a way that an efficient allocation of resources is made. For the price system acts to distil or encapsulate the information scattered among millions of individual economic actors in such a way that they are able to coordinate their efforts so that account is taken of all the information even though no single individual has access to all of it. For example, such circumstances as an increased demand for tin in one part of the economy, due to its utility in manufacturing some needed product, or the elimination of some source of tin, due, say, to an earthquake which destroys some mining operation, will affect the price of tin in such a way that users of tin will begin to economize – they will begin to use less of it, find alternatives to it, etc. In other words, a reallocation of resources toward their most efficient uses will take place even though no single individual knows of all the circumstances that led to there being a need for a reallocation – each individual needs know only that prices have changed, and this leads all individuals to act in a way that it might appear they could act only if directed by some central

authority. So what the socialist thinks can be done only by means of central planning of the economy in fact *cannot* be done in that fashion; but it can and is done through the activities of individuals responding to the signals of prices generated in a competitive market. (We thus have an instance of the operation of Adam Smith's "invisible hand.")

We might say that the decentralized, dispersed character of economic information in a market economy parallels the distributed, decentralized character of mental representations in a connectionist network, and that the unplanned and unplannable results of the economic process parallel the evolutionarily-arrived-at character of a highly developed connectionist mind. The lesson derived from philosophical psychology for social thought would be: you cannot plan an economic system from the top down, socialist style, any more than you can create artificial intelligence by fiat, in the fashion of AI approaches inspired by the classical symbolic processing paradigm.

The "tacit" component in economic knowledge just mentioned brings us to another respect in which an interesting connection exists between Hayek's philosophy of mind and his social philosophy. We saw earlier how Hayek takes the rules that govern intentionality and reason to be largely a result of biological and cultural evolution; but his primary application of this idea is in fact to *moral* rules, and to a defense of moral conservatism.

Hayek (1997) gives the label "constructivism" to the assumption, tacitly made in his view by socialism and radical movements in general, that basic moral and social institutions are, can be, and/or should be consciously and rationally designed for a particular purpose – the implication being that if we are, for whatever reason, unhappy with existing institutions, we can always tear them down and design new ones to replace them at will. The problem with this assumption, in his view, is that it assumes that reason, and the mind as a whole, is fully developed in some sense *apart from* and *prior to* social institutions, when in fact, for reasons we've seen, it *evolves with* such institutions and can never completely "step outside" of them.

It also assumes a false dichotomy between what is "natural," that is, not a result of human action (such as such unalterable facts of human biology as the drives to eat, sleep, and reproduce), and what is "artificial" or the result of deliberate human design (such as works of art and literature, machines, buildings, and so forth), and assumes that anything that doesn't fall into the first category – such as fundamental moral and social institutions – must fall into the second. But there is a *third* category, namely what the Scottish Enlightenment thinker Adam Ferguson called "the results of human action, but not of human design," structures that would not have existed had human beings not interacted with one another in the complex ways they do, but which nevertheless were not – *and could not have been* – deliberately intended or designed. The most obvious example would be language: it would not exist but for human interaction, and yet it obviously wasn't and couldn't have been *designed*, since design

presupposes intelligence, purpose, and planning, and these could surely not have existed to any great degree prior to the development of language; rather, the development of language itself is what made *them* possible. Language is an example of what Hayek, following the great figures of the Scottish Enlightenment, called "spontaneous orders," complex law-governed systems or structures having the appearance of conscious design but which actually arise through blind, impersonal processes, usually *evolutionary* processes. In the natural world, crystals, galaxies, and animal species are among the most impressive examples. In the human realm, language is but one example – others are, in Hayek's view, fundamental legal and cultural institutions, the market economy, and moral traditions, including especially (but not exclusively) those traditions (such as respect for private property and contracts) which underlie the market economy.

Such institutions and traditions are in Hayek's view for the most part not (and could not have been) the products of deliberate design, as constructivists tend at least implicitly to suppose, but rather the result of the sort of cultural evolution described earlier. Such cultural evolution is best thought of as involving the natural selection of *traditions* (rules of conduct, moral and otherwise, mores and taboos) and proceeds roughly as follows: rules the observance of which enables groups following them best to adapt to their environments will preserve those groups and allow them to grow and prosper, and are thus preserved themselves and copied by other groups; while rules which are not adaptive will cause groups following them to shrink, become

impoverished, and even die out. Again, the function and benefit of such rules are generally not known, or even knowable, before the fact; nor are they typically the reasons for their adoption, and the real reasons may even be superstitious. This parallels the process of biological evolution in which a mutation comes about, not because an animal sees some use for it, but through purely random genetic processes, and is preserved if it *in fact*, in a way the animal could not have foreseen, allows the animal better to adapt to its environment.

The lesson Hayek draws from this is that it is folly to suppose that we can redesign basic moral and social institutions at will – just as it would be folly to suppose that we could redesign an animal species at will and do a better job than evolution has at adapting it to its environment. We simply lack the knowledge required to do so. For the same reason, it is naïve to suppose that traditional moral institutions ought to be rejected if not supported by arguments which pass the sort of muster applied to philosophical and scientific theories. That such institutions may typically be undefended by those who practice them, or defended only on superstitious grounds, is irrelevant. What *is* relevant is the *function* they serve, and the fact that they have survived as long as they do is prima facie evidence that they *do* serve an important function, even if we do not, or cannot, know what it is. (Compare the way in which organs that often appear vestigial, such as tonsils, turn out on further investigation – and often after thousands have had them removed! – to serve some health-enhancing function after all.) Not that traditional practices must forever remain immune from

criticism; but in Hayek's view, traditions which have survived the test of time get the benefit of the doubt, and the burden of proof is always on those who want to abandon them, not on those who want to conserve them. For by the time we do find out what function they have served – upon observing the results of their abandonment – it may be too late to reinstitute them (especially given that beneficial as many traditional constraints are, Hayek recognizes that they are often disliked). (The catastrophe that befell those countries that turned to communism and abandoned adherence to the moral rules underpinning the market order – private property, respect for individual autonomy, and the like – is only the most dramatic of the sort of negative results of abandoning tradition Hayek has in mind; the unforeseen and unintended consequences of the so-called "sexual revolution" and the rise of the bureaucratic welfare state are other, more controversial, examples.)

Hayek's conception of the mind as inevitably governed by tacit rules and assumptions which ultimately can have no explicit explanation or justification (think in the moral realm of the way in which we ultimately just find certain things morally offensive and others as morally praiseworthy, arguably without being able satisfactorily to justify these fundamental intuitions) thus leads to a defense of a kind of *libertarian conservatism*. Society, like the mind, is too complex a system to try to construct or reconstruct wholesale, and is thus in general best left alone to develop as it will; and like the mind, as it develops, it will take on general characteristics which we do not and cannot fully understand or rationalize but without which it could not function

properly. Human society, like the human mind, is not something we can step outside of and remake from the top down (or bottom up); it makes us more than we make it.[13] And human progress, like the mind's evolution, is not something that we can, on the large scale anyway, bring about ourselves; it comes about blindly, in jerks and fits and with occasional setbacks, but arguably with an overall forward trajectory, as cultural evolution and the market process ratchet things up by producing and preserving moral, cultural, intellectual, and technological changes which are beneficial to the species and weeding out those which are not.[14]

Of course, this takes us far afield from the problems with which we have been mainly concerned in this essay, and it is not my purpose here to discuss in detail or defend Hayek's social and political philosophy.[15] We see in it, though, an extension of the theme that runs throughout Hayek's work in the philosophy of mind, namely the way in which the human condition is characterized ultimately by permanent, but too often unnoticed, limitations on our knowledge, and that the solution to many of the problems that plague us – whether they be problems about the mind's place in the

---

[13] There thus seems to be an instructive parallel between full-blown socialist attempts to remake social institutions from the ground up, and the eliminative materialist project of casting off "folk psychology" and redescribing human nature entirely in terms of concepts derived from neuroscience or cognitive science. Hayekian considerations imply that the latter project is as destined for failure as the first. In any case, Hayek would no doubt have looked on the latter with the same mixture of amusement and horror with which he looked on the former.

[14] This link between a conception of human knowledge as severely limited and a more or less conservative approach to politics and society is also to be seen not only in such thinkers as Edmund Burke and Michael Oakeshott, but also, perhaps most starkly, in the epistemological and political skepticism of Hume (whom Antony Flew (1986, pp. 172-175) regards as the true father of political conservatism, rather than Burke).

[15] For discussion and defense of an important part of it, see my "Hayek on Social Justice: Reply to Lukes and Johnston" (1997) and "Hayek, Social Justice, and the Market: Reply to Johnston" (1998b).

natural world or man's place in the social world – is to *recognize* these limitations, recognize that sometimes a problem only seems real, not because we know certain things, but because we erroneously *think* we do.

We have come full circle. The mind-body problem – especially the qualia problem, but also to some extent the problem of intentionality – has for centuries plagued philosophy and appeared to many theorists to pose an insurmountable barrier to the otherwise seemingly unstoppable advance of human knowledge. Ironically, that barrier itself is an illusion fostered by a radical *lack* of knowledge on our part: it disappears when we recognize how little we really know, about the intrinsic natures of the physical and mental worlds, and about the principles which ultimately govern the latter. Socrates' admonition to all philosophers has been borne out: the beginning of knowledge is the recognition of our ignorance.

Bibliography

Agonito, Rosemary (1975), 'Hayek Revisited: Mind as the Process of Classification', Behaviorism 3 (2), pp. 162-171.

Armstrong, D.M. (1961), Perception and the Physical World (London: Routledge and Kegan Paul).

Armstrong, D.M. (1968), A Materialist Theory of the Mind (London: Routledge and Kegan Paul).

Austin, J.L. (1962), Sense and Sensibilia (New York: Oxford University Press).

Ayer, A.J. (1940), The Foundations of Empirical Knowledge (London: Macmillan).

Barrow, John D. (1992), Pi in the Sky: Counting, Thinking, and Being (Oxford: Clarendon Press).

Baumgartner, Peter and Sabine Payr (1995), Speaking Minds: Interviews with Twenty Eminent Cognitive Scientists (Princeton: Princeton University Press).

Block, Ned (1978), 'Troubles With Functionalism' in Minnesota Studies in the Philosophy of Science, Vol. IX ed. C. Wade Savage (Minneapolis: University of Minnesota Press).

Block, Ned (1994), 'Consciousness' in A Companion to the Philosophy of Mind ed. Samuel Guttenplan (Oxford: Basil Blackwell).

Braithwaite, R.B. (1964), 'An Empiricist's View of the Nature of Religious Belief' in John Hick, ed. The Existence of God (New York: Macmillan).

Brentano, Franz (1995), Psychology from an Empirical Standpoint (London: Routledge).

Budd, Malcolm (1989), Wittgenstein's Philosophy of Psychology (London: Routledge).

Carroll, Lewis (1977), Symbolic Logic (New York: Clarkson N. Potter, Inc.)

Chalmers, David (1995), 'Facing up to the problem of consciousness', Journal of Consciousness Studies, 2 (3), pp. 200-19.

Chalmers, David (1996), The Conscious Mind (New York: Oxford University Press).

Churchland, Paul M. (1981), 'Eliminative Materialism and the Propositional Attitudes', Journal of Philosophy 78 (2).

Churchland, Paul M. (1988), Matter and Consciousness, Revised Edition (Cambridge, MA: The MIT Press).

Churchland, Paul M. (1989a), 'Reduction, Qualia, and the Direct Introspection of Brain States' in A Neurocomputational Perspective (Cambridge, MA: The MIT Press).

Churchland, Paul M. (1989b), 'Knowing Qualia: A Reply to Jackson' in A Neurocomputational Perspective (Cambridge, MA: The MIT Press).

Dempsey, Gary T. (1996), 'Hayek's *Terra Incognita* of the Mind', The Southern Journal of Philosophy 34.

Dennett, Daniel (1991), Consciousness Explained (Boston, MA: Little, Brown, and Co.).

Dennett, Daniel (1993), 'Quining Qualia', in Readings in Philosophy and Cognitive Science ed. Alvin I. Goldman (Cambridge, MA: The MIT Press).

Dretske, Fred (1995), Naturalizing the Mind (Cambridge, MA: The MIT Press).

Dreyfus, Hubert L. and Stuart E. Dreyfus (1990), 'Making a Mind Versus Modelling the Brain: Artificial Intelligence Back at a Branchpoint' in The Philosophy of Artificial Intelligence ed. Margaret A. Boden (New York: Oxford University Press).

Edelman, Gerald (1982), 'Through a Computer Darkly: Group Selection and Higher Brain Function', Bulletin – The American Academy of Arts and Sciences 36 (1).

Edelman, Gerald (1987), Neural Darwinism (New York: Basic Books).

Feigl, Herbert (1967), The 'Mental' and the 'Physical': The Essay and a Postscript (Minneapolis: University of Minnesota Press).

Feser, Edward (1997), 'Hayek on Social Justice: Reply to Lukes and Johnston', Critical Review 11 (4), pp. 581-606.

Feser, E. (1998a), 'Can Phenomenal Qualities Exist Unperceived?', Journal of Consciousness Studies 5 (4), pp. 405-414.

Feser, Edward (1998b), 'Hayek, Social Justice, and the Market: Reply to Johnston', Critical Review 12 (3), pp. 269-281.

Flanagan, Owen (1992), Consciousness Reconsidered (Cambridge, MA: The MIT Press).

Fleetwood, Steve (1995), Hayek's Political Economy (London: Routledge).

Flew, Antony (1984), God, Freedom, and Immortality (Buffalo: Prometheus Books).

Flew, Antony (1986), David Hume: Philosopher of Moral Science (Oxford: Basil Blackwell).

Fodor, Jerry A. (1981), 'Three Cheers for Propositional Attitudes' in Representations (Cambridge, MA: The MIT Press).

Fodor, Jerry A. (1987), Psychosemantics (Cambridge, MA: The MIT Press).

Fodor, Jerry A. (1994), 'Fodor, Jerry A.' in A Companion to the Philosophy of Mind ed. Samuel Guttenplan (Oxford: Basil Blackwell).

Forsyth, Murray (1988), 'Hayek's Bizarre Liberalism: A Critique', Political Studies XXXVI, pp. 235-250.

Foster, John (1982), The Case for Idealism (London: Routledge and Kegan Paul).

Foster, John (1991), The Immaterial Self (London: Routledge).

Frege, Gottlob (1967), 'The Thought: A Logical Inquiry' in Philosophical Logic ed. P.F. Strawson (Oxford: Oxford University Press).

Fuster, Joaquin (1995), Memory in the Cerebral Cortex: An Empirical Approach to Neural Networks in the Human and Nonhuman Primate (Cambridge, MA: The MIT Press).

Glock, Hans-Johann (1996), A Wittgenstein Dictionary (Oxford: Blackwell).

Gray, John (1993), Post-liberalism: Studies in Political Thought (London: Routledge).

Gray, John (1998), Hayek on Liberty, Third Edition (London: Routledge).

Grayling, A.C. (1996), Russell (New York: Oxford University Press).

Grice, H.P. (1988), 'The Causal Theory of Perception' in Perceptual Knowledge ed. Jonathan Dancy (Oxford: Oxford University Press).

Grim, Patrick (1984), 'There Is No Set of All Truths', Analysis 44, pp. 206-208.

Hamlyn, D.W. (1954), 'Review of *The Sensory Order*', Mind 252, pp. 560-562.

Hardin, C.L. (1997), 'Reinverting the Spectrum', in Readings on Color, Volume 1: The Philosophy of Color ed. Alex Byrne and David R. Hilbert (Cambridge, MA: The MIT Press).

Hart, William D. (1988), The Engines of the Soul (New York: Cambridge University Press).

Hart, William D. (1994), 'Dualism' in A Companion to the Philosophy of Mind ed. Samuel Guttenplan (Oxford: Basil Blackwell).

Hayek, F.A. (1944), The Road to Serfdom (Chicago: The University of Chicago Press).

Hayek, F.A. (1952), The Sensory Order (Chicago: University of Chicago Press).

Hayek, F.A. (1967), 'Rules, Perception, and Intelligibility' in Studies in Philosophy, Politics, and Economics (London: Routledge and Kegan Paul).

Hayek, F.A. (1973), Law, Legislation, and Liberty, Volume I: Rules and Order (Chicago: The University of Chicago Press).

Hayek, F.A. (1978), 'The Primacy of the Abstract', in New Studies in Philosophy, Politics, Economics and the History of Ideas (London: Routledge and Kegan Paul).

Hayek, F.A. (1979), The Counter-Revolution of Science (Indianapolis: Liberty Press).

Hayek, F.A. (1982), 'The Sensory Order After 25 Years' in Cognition and the Symbolic Processes, Vol. II ed. W.B. Weimer and D.S. Palermo (Hillsdale, N.J.: Lawrence Erlbaum).

Hayek, F.A. (1988) The Fatal Conceit: The Errors of Socialism (Chicago: University of Chicago Press).

Hayek, F.A. (1994) Hayek on Hayek (Chicago: Uniersity of Chicago Press).

Hayek, F.A. (1997a), 'The Use of Knowledge in Society' in The Libertarian Reader ed. David Boaz (New York: The Free Press).

Hayek, F.A. (1997b), 'The Errors of Constructivism' in Conservatism: An Anthology of Social and Political Thought from David Hume to the Present ed. Jerry Z. Muller (Princeton: Princeton University Press).

Hayek, F.A. and Walter B. Weimer (1982), 'Weimer-Hayek Discussion' in Cognition and the Symbolic Processes, Vol. II ed. W.B. Weimer and D.S. Palermo (Hillsdale, N.J.: Lawrence Erlbaum).

Hebb, D.O. (1949), The Organization of Behavior (New York: Wiley).

Hofstadter, Douglas R. (1979), Gödel, Escher, Bach: An Eternal Golden Braid (New York: Vintage Books).

Hofstadter, Douglas R. (1981), "'Reflections" on "What is it like to be a bat?"' in The Mind's I ed. Douglas R. Hofstadter and Daniel C. Dennett (New York: Bantam Books).

Holland, John H. (1992a), 'Complex Adaptive Systems', Daedalus 121 (1), pp. 17-30.

Holland, John H. (1992b), 'Genetic Algorithms', Scientific American 267 (1), pp. 66-72.

Holland, John H. (1995), Hidden Order (New York: Addison-Wesley).

Horgan, John (1997), The End of Science (New York: Broadway Books).

Jackson, Frank (1982), 'Epiphenomenal Qualia', Philosophical Quarterly, 32, pp. 127-36.

Jackson, Frank (1991), 'What Mary Didn't Know' in The Nature of Mind ed. David Rosenthal (New York: Oxford University Press).

Kaplan, David (1990), 'Quantifying In' in The Philosophy of Language, Second Edition ed. A.P. Martinich (New York: Oxford University Press).

Kirk, Robert (1994), Raw Feeling (Oxford: Clarendon Press).

Kripke, Saul (1971), 'Naming and Necessity' in Semantics of Natural Language ed. Donald Davidson and Gilbert Harman (Dordrecht: Reidel).

Kuhn, Thomas S. (1970), The Structure of Scientific Revolutions, Second Edition (Chicago: University of Chicago Press).

Levine, Joseph (1993), 'On Leaving Out What It's Like' in Consciousness ed. Martin Davies and Glyn W. Humphreys (Oxford: Blackwell).

Levine, Joseph (1995), 'Qualia: Intrinsic, Relational, or What?' in Conscious Experience ed. Thomas Metzinger (Schoningh: Imprint Academic).

Lewis, David (1991), 'Postscript to "Mad Pain and Martian Pain: "Knowing What It's Like"' in The Nature of Mind ed. David Rosenthal (New York: Oxford University Press).

Lockwood, Michael (1981), 'What *Was* Russell's Neutral Monism?' in Midwest Studies in Philosophy Vol. VI: The Foundations of Analytic Philosophy ed. Peter A. French, Theodore E. Uehling, Jr., and Howard K. Wettstein (Minneapolis: University of Minnesota Press).

Lockwood, Michael (1989), Mind, Brain, and the Quantum (Oxford: Basil Blackwell).

Lockwood, Michael (1993), 'The Grain Problem' in Objections to Physicalism ed. Howard Robinson (Oxford: Clarendon Press).

Lockwood, Michael (1998), 'Unsensed Phenomenal Qualities: A Defence', Journal of Consciousness Studies 5 (4), pp. 415-418.

Lucas, J.R. (1961), 'Minds, Machines, and Gödel', Philosophy 36, pp. 120-124.

Mackie, J.L. (1976), Problems from Locke (Oxford: Clarendon Press).

Mackie, J.L. (1982), The Miracle of Theism (Oxford: Clarendon Press).

Martens, David B. (1992), 'Knowledge by acquaintance/by description' in A Companion to Epistemology ed. Jonathan Dancy and Ernest Sosa (Oxford: Basil Blackwell).

Maxwell, Grover (1978), 'Rigid Designators and Mind-Brain Identity', in Minnesota Studies in the Philosophy of Science, Vol. IX ed. C. Wade Savage (Minneapolis: University of Minnesota Press).

McDowell, John (1986), 'Singular Thought and the Extent of Inner Space' in Subject, Thought, and Context eds. J. McDowell and P. Pettit (Oxford: Oxford University Press).

McGinn, Colin (1991), The Problem of Consciousness (Oxford: Basil Blackwell).

Meehl, P.E. (1966), 'The Compleat Autocerebroscopist: A Thought-Experiment on Professor Feigl's Mind-Body Identity Thesis' in Mind, Matter, and Method: Essays in Philosophy and Science in Honor of Herbert Feigl ed. P.K. Feyerabend and G. Maxwell (Minneapolis: University of Minnesota Press).

Miller, Eugene F. (1979) 'The Cognitive Basis of Hayek's Political Thought' in Liberty and the Rule of Law ed. Robert Cunningham (College Station: Texas A and M University Press).

Miller, Mark S. (1996), 'Learning Curve', Reason, 28 (7).

Monk, Ray (1990), Ludwig Wittgenstein: The Duty of Genius (New York: The Free Press).

Moore, G.E. (1962), Some Main Problems of Philosophy (New York: Collier Books).

Nagel, Thomas (1974), 'What is it like to be a bat?', Philosophical Review, 4, pp. 435-50.

Nagel, Thomas (1995), Other Minds (New York: Oxford University Press).

Penrose, Roger (1989), The Emperor's New Mind (New York: Oxford University Press).

Penrose, Roger (1994), Shadows of the Mind (New York: Oxford Univesity Press).

Phillips, D.Z. (1970), Death and Immortality (London: Macmillan).

Phillips, D.Z. (1976), Religion Without Explanation (Oxford: Basil Blackwell).

Place, U.T. (1956), 'Is Consciousness a Brain Process?', British Journal of Psychology, 47, pp. 44-50.

Popper, Karl and John Eccles (1977), The Self and Its Brain (London: Routledge and Kegan Paul).

Putnam, Hilary (1975), 'The Meaning of "Meaning"' in Minnesota Studies in the Philosophy of Science, Volume 7: Language, Mind, and Knowledge ed. K. Gunderson (Minneapolis: University of Minnesota Press).

Putnam, Hilary (1981), Reason, Truth, and History (New York: Cambridge University Press).

Putnam, Hilary (1991), 'The Nature of Mental States' in The Nature of Mind ed. David Rosenthal (New York: Oxford University Press).

Putnam, Hilary (1994), Words and Life (Cambridge, MA: Harvard University Press).

Quine, W.V. (1950), 'Two Dogmas of Empiricism', The Philosophical Review 60, pp. 20-43.

Robinson, Howard (1994), Perception (London: Routledge).

Rucker, Rudy (1982), Infinity and the Mind (Boston: Birkhauser).

Russell, Bertrand (1945), A History of Western Philosophy (New York: Simon and Schuster).

Russell, Bertrand (1948), Human Knowledge (New York: Simon and Schuster).

Russell, Bertrand (1954), The Analysis of Matter (New York: Dover).

Russell, Bertrand (1956), 'Mind and Matter' in Portraits from Memory (London: George Allen and Unwin).

Russell, Bertrand (1959), My Philosophical Development (London: George Allen and Unwin).

Russell, Bertrand (1978), The Analysis of Mind (New York: Humanities Press).

Russell, Bertrand (1988), The Problems of Philosophy (Buffalo, NY: Prometheus Books).

Ryle, Gilbert (1949), The Concept of Mind (New York: Barnes and Noble).

Schlick, Moritz (1985), General Theory of Knowledge (La Salle, Ill.: Open Court).

Scruton, Roger (1995), Modern Philosophy: An Introduction and Survey (New York: Penguin Books).

Searle, John R. (1980), 'Minds, Brains, and Programs', Behavioral and Brain Sciences III (3), pp. 417-424.

Searle, John (1983), Intentionality (Cambridge: Cambridge University Press).

Searle, John (1984), Minds, Brains, and Science (Cambridge, MA: Harvard University Press).

Searle, John (1992), The Rediscovery of the Mind (Cambridge, MA: The MIT Press).

Searle, John R. (1997), The Mystery of Consciousness (New York: The New York Review of Books).

Sellars, Wilfrid (1956), 'Empiricism and the Philosophy of Mind' in Minnesota Studies in the Philosophy of Science, Volume I: The Foundations of Science and the Concepts of Psychology and Psychoanalysis ed. Herbert Feigl and Michael Scriven (Minneapolis: University of Minnesota Press).

Smart, J.J.C. (1959), 'Sensations and Brain Processes', The Philosophical Review, 68, pp.141-56.

Smith, Barry (1994), Austrian Philosophy: The Legacy of Franz Brentano (La Salle, Ill.: Open Court).

Smith, Barry (1996), 'The Connectionist Mind: A Study of Hayekian Psychology' in Hayek the Economist and Social Philosopher: A Critical Retrospect ed. S. Frowen (London: Macmillan).

Snowden, Paul (1994), 'Neutral Monism' in The Oxford Companion to Philosophy ed. Ted Honderich (New York: Oxford University Press).

Sprott, W.J.H. (1954), 'Review of *The Sensory Order*', Philosophy 109, pp. 183-185.

Strawson, Galen (1994), Mental Reality (Cambridge, MA: The MIT Press).

Strawson, Galen (forthcoming), 'Realistic Monism' in Chomsky and his Critics ed. L. Antony and N. Hornstein (Oxford: Blackwell).

Swinburne, Richard (1986), The Evolution of the Soul (Oxford: Clarendon Press).

Tipler, Frank J. (1994), The Physics of Immortality (New York: Doubleday).

Tye, Michael (1995), Ten Problems of Consciousness (Cambridge, MA: The MIT Press).

Van Gulick, Robert (1993), 'Understanding the Phenomenal Mind: Are We All just Armadillos?' in Consciousness ed. Martin Davies and Glyn W. Humphreys (Oxford: Blackwell).

de Vries, Robert P. (1994), 'The Place of Hayek's Theory of Mind and Perception in the History of Philosophy and Psychology' in Hayek, Co-ordination, and Evolution ed. Jack Birner and Rudy van Zijp (London: Routledge).

Weimer, Walter B. (1982), 'Hayek's Approach to the Problems of Complex Phenomena: An Introduction to the Theoretical Psychology of *The Sensory Order*', in Cognition and the Symbolic Processes, Vol. II ed. W.B. Weimer and D.S. Palermo (Hillsdale, N.J.: Lawrence Erlbaum).

Wittgenstein, Ludwig (1953), Philosophical Investigations (Oxford: Blackwell).

Wittgenstein, Ludwig (1967), Zettel (Oxford: Blackwell).

Wittgenstein, Ludwig (1969), On Certainty (Oxford: Blackwell).

Yablo, Stephen (1993), 'Is Conceivability a Guide to Possibility?', Philosophy and Phenomenological Research, LIII (1), pp. 1-42.