Minds, Machines and Qualia: A Theory of Consciousness

by

Christopher Williams Cowell

A.B. (Harvard University) 1992

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Philosophy

in the

GRADUATE DIVISION

of the

UNIVERSITY OF CALIFORNIA, BERKELEY

Committee in charge:

Professor John R. Searle, Chair
Professor Hans D. Sluga
Professor John F. Kihlstrom

Spring 2001

The dissertation of Christopher Williams Cowell is approved:

Chair Date

Date

Date

University of California, Berkeley

Spring 2001

Minds, Machines and Qualia: A Theory of Consciousness

by

Christopher Williams Cowell

Abstract

Minds, Machines and Qualia: A Theory of Consciousness

by

Christopher Williams Cowell

Doctor of Philosophy in Philosophy

University of California, Berkeley

Professor John R. Searle, Chair

It is clear that there is a problem of consciousness; it is less clear what that problem is. In chapter one I discuss what it might be, bracket off some especially intractable issues, and present two central questions. First, what is the nature of consciousness? Second, what are the prospects for producing consciousness in machines? I then look at various ways one might approach these questions.

Chapter two focuses on the nature of consciousness. My definition of consciousness centers on qualia, a concept that I discuss in detail. I show that consciousness can be thought of as having three aspects: intransitive creature consciousness, transitive creature consciousness, and state consciousness. The relations between these three facets are explored.

Chapter three expands on two issues raised in chapter two. First, I argue that qualia are present not just in sense perception but also in ordinary propositional thought.

Second, I contrast my definition of consciousness with other common definitions. I show that some of these reduce to qualia-based theories similar to mine, while others lead to problems severe enough to suggest that they should be abandoned.

Chapter four deals with machine consciousness by looking carefully at the Chinese room argument and at connectionist models of cognition. I support the Chinese room's conclusion that the running of a program is insufficient for producing consciousness, and I argue that the possibility of machine consciousness hinges on our ability to identify and replicate the causal mechanisms that produce it in humans.

My theory has implications for a variety of philosophical questions. These implications are explored in a somewhat speculative manner in chapter five. I discuss how consciousness affects our efforts to solve the mind/body problem; I assess possible strategies for recognizing consciousness in other systems; and I examine connections between consciousness and personhood, focusing on the issues of personal identity and personal rights.

Instead of developing a single line of thought or argument in this dissertation, I present a collection of ideas and arguments which, although not tightly integrated, do provide mutual support and contribute to a single overarching theory of consciousness.

Professor John R. Searle
Dissertation Committee Chair

I dedicate this work to my parents and brother, three of the best people I know.

Contents

List of Figures

Acknowledgments

First, the technical stuff. This dissertation was produced entirely with free software. Writing and editing were done mostly with the NEdit text editor on a home-built Celeron box running Mandrake Linux and the Enlightenment window manager. Some editing was also done with BBEdit Lite and Tom Bender's Tex-Edit Plus on a Macintosh Powerbook Duo. Typesetting was done with Donald Knuth's TEX, using Leslie Lamport's LATEX 2_ε macros, a lightly modified version of Karl Berry and Oren Patashnik's BIBTEX bibliography template, and a tweaked version of the University of California Thesis Class by Ethan V. Munson and Blaise B. Frederick. The figures were created within TEX's picture environment, and I used Markku Hihnala's kdvi for proofing the typeset copy. The typeface used throughout is Knuth's Computer Modern. I've included this information because I think it is important to recognize and support those who write useful freeware or shareware, especially when it is open source (as many of these packages are).

Second, and of course far more importantly, the people. My parents, Richard and Priscilla, and my brother Nicholas have supported me in countless ways not just during graduate school, but throughout my entire life. I consider myself *extremely* lucky to have grown up with people that I would have liked and respected enormously even if I hadn't

been related to them. They have shaped my values, my ways of thinking, and my world-view more than they probably realize; I love all three very much.

I also owe a huge debt of gratitude to Kinnari Shah for her unending support. She encouraged me when progress was slow, and she was quick to recognize and celebrate even the most minor of milestones along the way. She is due an extra dollop of thanks for taking on (requesting, even!) the mammoth job of copyediting this entire work—a duller task I cannot imagine. Many thanks and lots of love to Kinnari.

Of course I am extremely grateful to the members of my dissertation committee. Hans Sluga and John Kihlstrom influenced the shape and direction of this work with their very helpful comments. John Searle not only served as my committee chairman and main advisor, but has been in many ways my primary philosophical role model. I have learned from him an enormous number of philosophical ideas, approaches, and arguments, but even more important are the lessons I have picked up from him about how to present philosophical content in a straightforward, non-obfuscatory fashion. Throughout graduate school, my writing has been strongly influenced by his sterling maxim, "if you can't say it clearly, you don't understand it yourself." The philosophical community—come to think of it, the whole academy—would be better, more productive, and certainly more accessible if more scholars took his words to heart. I hope my writing in this project reflects my appreciation of this principle; I have deliberately chosen a simple, unsophisticated style in an effort to allow the reader to concentrate on evaluating my arguments without being burdened by having to untangle and decipher convoluted prose. I have tried very hard to imitate his clarity of expression, his frequent use of vivid examples, and his refreshingly

common-sensical approach to philosophy in general.

Other philosophers who have had an especially strong influence on my thoughts and my way of communicating these thoughts include Berkeley faculty members Janet Broughton, Daniel Warren, and Richard Wollheim. I also include in this group Eric Lormand and Steve Yablo from the Michigan philosophy department while I was there in 1993–95, and Robert Nozick, Sarah Patterson, and Susan Sauvé from Harvard's department in the early 1990s. I have had the enormous privilege of studying in all three of these departments, as well spending a year doing dissertation work in the Princeton philosophy department. I have always believed that exposure to a variety of approaches to a given problem can only help to sharpen one's own approach and ideas, and I think this is especially true when the problems are as nebulous as those typically found in philosophy.

I should also recognize and thank Scott Baker for being a willing resource for all matters biological, and Kirsten Holmboe for instructing me on some of the rudiments of human neurophysiology. Eddie Cushman and Peter Hanks provided invaluable assistance as I worked through some thorny issues in chapters one and two. Arpy Khatchirian, Jennifer Johnson, and Jessica Gelber were always willing to discuss philosophical and non-philosophical matters, and were helpful in reminding me that occasional non-philosophical interruptions enrich rather than dilute one's intellectual life. Finally, I want to thank and apologize to the hundreds of Berkeley undergraduates who have allowed me to abuse my position as their teaching assistant by forcing them to listen to my views on consciousness and the role it plays in our lives. These students often pose comments and questions that are far more direct, relevant, and troubling than many of the points raised in professional

philosophy journals.

It is a commonplace in philosophical writing to share credit for the ideas one presents while claiming full responsibility for any errors, inconsistencies, or incoherence. I will continue in this tradition, but I do so more in the spirit of acknowledging these problems than of apologizing for them. The boundaries of our knowledge can only be expanded if we are willing to take intellectual risks; we must have the courage to place new concepts or arguments—outlandish though they may seem—in front of other thinkers and then gauge how well they withstand careful scrutiny. So while I would be very surprised if *all* of the claims in this work are in fact true, this fact does not worry or embarrass me. I hope that *enough* of this work is true to illuminate some murky issues and to help us see possible ways out of a few of the extremely challenging problems found in contemporary philosophy of mind.

Consciousness. *The having of perceptions, thoughts, and feelings; awareness. The term is impossible to define except in terms that are unintelligible without a grasp of what consciousness means.... Consciousness is a fascinating but elusive phenomenon; it is impossible to specify what it is, what it does, or why it evolved. Nothing worth reading has been written on it.*

Stuart Sutherland, *The International Dictionary of Psychology* [105]

Chapter 1

The problem of consciousness

> *Consciousness is a subject about which there is very little consensus, even as to what the problem is.*
>
> Francis Crick [21, p. xii]

1.1 What is the problem of consciousness?

The topic of consciousness, after a long period of neglect in the philosophical and psychological literature, is back with a vengeance. Behaviorism is dead, and its demise has freed us to speak of consciousness without using the hushed, somewhat sheepish tones that are still de rigueur for discussions of "mystical" topics such as Bergsonian *élan vital* or even Cartesian dualism. Scholars seem to approve of and enjoy this new freedom: the volume of writing on consciousness that has appeared in the last decade or so is staggering, as is the variety of approaches and directions taken. Academics from a huge range of disciplines and backgrounds have begun to try their hands at saying something about consciousness, with predictably mixed results. Some of this writing has been quite helpful, some less so, and

much has been—to put it charitably—just plain confused. But almost every person who has tackled the subject agrees that there exists a definite and difficult *problem of consciousness*. What they fail to agree on is what this problem is exactly. The range of questions that different authors have in mind when referring to the problem of consciousness is quite broad, and these questions range from general to very specific. A list of just a few of these conveys a sense of how little consensus there is about what the problem of consciousness amounts to:

- What is consciousness? How should we define it?
- Is it a thing, a process, an event, or something else? Does it fit neatly into any preëxisting ontological category, or is it *sui generis*?
- Are there different kinds of consciousness?
- Who has it? Do all humans? Do computers? Animals? Plants?
- Do some people have more of it than others?
- Can it exist apart from a human brain? Can it exist without any physical substratum at all?
- How can we test for its presence in a system? If something acts intelligently, does that mean it is conscious? Why do we feel so sure that other people have it but that rocks and rosebushes don't?
- What methods and tools can be used for researching consciousness?
- Can two people share the same consciousness, either concurrently or serially?

- Is there a seat of consciousness? What might this mean?
- How does the brain produce it? What parts of the brain are responsible? What is the neural correlate of consciousness?
- How does it interact with the body?
- Can it be created and destroyed, or is it eternal?
- What does it do for us? Are there things we can do with it that we could not do without it?
- Are there unconscious or subconscious states? If so, what is the relation between these states and conscious states?
- Why does an apple taste like an apple and not like something else? Why does an apple taste like anything at all?
- Is your green the same as my green? Is your pain the same as my pain? Does an apple taste the same way to you as it does to me? Can this be tested? Are there even answers to such questions?
- How did we come to have consciousness? Did it evolve or was it accidental? Which of its properties or traits, if any, preserved and promoted it during the process of natural selection?
- What is the relation between consciousness and free will? Is either a necessary and/or sufficient condition of the other?

- What is the relation between consciousness and personhood? What does it mean to say that the consciousness in a body at one time is the same as the consciousness in the same body at a different time? Should we ever say that?

- What are the ethical implications of consciousness? Do all conscious systems have certain rights or responsibilities?

These questions are all worth addressing at great length, and each can be legitimately considered to make up part of the overall problem of consciousness. But I want to cut the problem of consciousness down to a more manageable size. My aim here is to focus on what I consider to be the two most interesting—and probably among the more tractable—questions concerning consciousness. First, *what is the nature of consciousness*? That is, how ought we to define it, and what can we say about its metaphysical status? What is it that all conscious systems have in common? Second, *what are the prospects for producing consciousness in machines, such as computers or robots*? Whenever I refer to "the problem of consciousness" from here on out, I will have these two fundamental questions in mind. However, as is the case with many issues in philosophy, the boundaries of all of the questions in the list above are both amorphous and porous. This means that my task of isolating and approaching just two of the questions in this large cluster of consciousness-related problems is bound to be a difficult one. Nevertheless, limits to any project do have to be drawn, and I will do my best to focus exclusively on these two central questions through chapter four. These issues are not self-contained, and work done on any one question from the list will likely have implications for many of the other questions. With this in mind, chapter five will be dedicated to an exploration of the implications that my answers to these two central

questions might have for some of the other consciousness-related questions I have listed.

Special mention needs to be made of three questions that I am deliberately leaving out of my formulation of the problem of consciousness. First, I will largely ignore the question of whether consciousness has any functions, and if so, what they might be. Considering the enormous evolutionary expense of developing consciousness over the millennia and the incredible pervasiveness of consciousness in virtually all aspects of everyday life, it would be truly astounding if it had no causal powers and served no function. The view that consciousness is the foam upon the waves rather than the waves themselves—a common (though to my ear, imperfect) metaphor used to suggest that the consciousness attached to a person merely goes along for the ride without actually influencing that person's behavior in any way—is called "epiphenomenalism," and is still considered a very live (if startlingly odd) option among many philosophers of mind. While I do admit to leaning slightly toward epiphenomenalism when forced to take a stand on the issue, I will not argue here either for this position or for the opposing view that consciousness does have a function or functions. This topic is not only fascinating but also vitally important, considering that epiphenomenalism has direct and devastating implications for freedom of the will. However, it is simply too large and complex an issue for me to address adequately here.

Second, I will not tackle the topic of the self in any significant detail. Consciousness is often described as requiring a subject—sometimes called a "self" or "ego"—that bears conscious states. While I agree that the notion of there being some entity which *has* or *experiences* conscious states is intuitively very appealing, there are a number of problems that arise with such a view. To take just one example, it would seem to require that

the self is independent of consciousness in some sense, and is capable of existing with no conscious states at all or perhaps even when disembodied. This raises all sorts of questions and worries about the exact ontological status of such a self, which often lead in turn to vague and unsatisfying claims about the self being "spiritual" or "soul-like."[1] While I do think that a full understanding of consciousness will ultimately require some resolution to the question of whether a self is required for consciousness to take root in, and if so, what sort of thing that self might be, those are not questions that my project demands answers to. I think I can make helpful observations about what consciousness is without necessarily taking a stand on the relation between consciousness and a putative self. This issue will arise somewhat in my discussion of personal identity in chapter five, but even there I will make only limited claims about the self.

Owen Flanagan poses particularly graphically the third important question that I am deliberately neglecting:

> How does conscious experience arise from a system of 100 billion well-connected neurons set in a soupy mixture of neurochemicals in which the only things that happen are activations of sets of neurons at different rates and changes in the composition of the soup? [33, p. 51]

How indeed does the human brain produce consciousness? How does an enormous mass of rather simply constructed neurons produce the taste of sugar, the feeling of being in love, or the giddy, happy fear one feels while riding a roller coaster? No amount of peering into or prodding a brain suggests an answer. Brains are objective, physical objects that we can interact with directly; virtually any sort of test can be run on a brain that can be run on

[1] This is not to denigrate the contributions that religion can make toward solving philosophical problems, or vice versa. I firmly believe that interaction between philosophers and theologians can be fruitful and in fact quite valuable, even if the two sides are not always sure how exactly to engage each other.

any other physical object. How much does it weigh? What color is it? How does it behave in a wind tunnel? But this is not true of ephemeral, subjective consciousness. We assume that everyone who has walked the earth has been accompanied by some form or other of consciousness, but as the still-unsolved problem of other minds reminds us, we have no concrete proof of these other instances of consciousness. This is because we lack any way of interacting directly with consciousnesses other than our own. Consciousness cannot be seen, and unless it is our own, it cannot be directly accessed or measured in any way. In short, it seems to be an entirely different sort of thing than a brain. The question of how a brain gives rise to consciousness is at once obvious and among the most difficult questions in all of the philosophy of mind. It is so difficult that it has been dubbed the *hard problem* in order to distinguish it from "easier" consciousness-related problems like understanding how memories are laid down, how learning occurs, or how emotion affects behavior.[2] The boundaries of the hard problem, like so many issues in consciousness, are somewhat nebulous. Some philosophers include the problem of mental-to-physical causation under its rubric, while others add the question of why brain processes produce the qualia they do, and not other qualia. But most use the phrase to describe the problem in its barest form: how does the physical brain cause qualitative properties of mental states? This problem, fascinating though it is, lies well beyond the scope of this dissertation. I fear that even the most superficial of answers to the hard problem would draw on a number of disciplines outside of philosophy (including, most prominently, psychology and neurobiology), and this effectively excludes me from competently addressing the issue.

[2]The origins of this name are obscure. It may have been coined by Galen Strawson, though it was not in common usage until popularized by David Chalmers at the 1994 *Toward a Science of Consciousness* conference at the University of Arizona.

There is another, purely philosophical, reason to avoid the hard problem. It is sometimes said (accurately, I think) that the hard problem points to an "explanatory gap," meaning that even a fully developed explanation of the workings of the brain would give us no insight into how those mechanical processes can produce consciousness. There seems to be a prominent and troubling gap that needs to be filled. But making even a preliminary effort at bridging this gap would require a comprehensive analysis of the nature of scientific explanation, not to mention an account of how explanations can span ontological categories in general. And this is much too large and thorny a topic for me to tackle here. But the two main questions concerning consciousness that I will address—what is consciousness and can it be produced artificially—are interesting in their own right, and should prove challenging enough.

1.2 Approaching the problem

> *I believe that when the neural basis of consciousness is thoroughly understood this knowledge will suggest answers to two major questions: What is the general nature of consciousness, so that we can talk sensibly about the nature of consciousness in other animals, and also in man-made machines, such as computers? What advantage does consciousness give an organism, so that we can see why it has evolved?*
>
> Francis Crick [21, p. 252]

After explicitly presenting these questions, Crick makes frustratingly little progress toward answering either one in his recent book on consciousness. I believe his failure stems from a reverse form of the problem I mentioned above: he unnecessarily limits his investigation by employing only a purely neurobiological approach. He justifies the inclusion of neuroscience among the class of disciplines that can contribute to our understanding of

the brain by noting, correctly, that "it is not sensible to tackle a very difficult problem with one hand tied behind one's back," but he then ironically ignores his own advice by neglecting philosophical considerations wholesale [21, p. 18]. While neurobiology will certainly play a crucial role in a completed theory of the mind and its relation to the brain, I believe that it is not the best tool to wield against the problems he and I are both interested in. In order to make further progress on questions concerning the nature and function of consciousness, we must turn away from the data-rich but ultimately unilluminating (for this problem at least—I don't mean to disparage science in general) approach of admittedly philosophically-minded neurobiologists like Crick, Gerald Edelman [32], and Christof Koch [23]; psychologists like Bernard Baars [6]; or medical researchers like anesthesiologist Stuart Hameroff [42, 43]; and turn instead to the tools used by neurobiologically aware philosophers: tools such as introspection and rigorous analysis. The purpose of this dissertation is precisely that; I want to use the instruments and techniques of analytic philosophy to take some preliminary steps toward answering these two questions.

I do not want to imply that science is useless when it comes to explaining consciousness. Cognitive psychology has increased enormously our understanding of human behavior and the high-level workings of the mind, while neurobiology has told us a great deal about the low-level architecture that supports memory, perception, learning, emotion, and all of the other components of mental life. Much of the most successful consciousness research that has been done in the last few years has involved the collection of functional magnetic resonance imaging (FMRI) and electroencephalogram (EEG) data, and as such has had a distinctly scientific rather than philosophical flavor. This research—primarily by

neurobiologists and psychologists—has helped us make great strides in understanding the roles played by particular brain structures in producing certain mental phenomena. We have unraveled to some extent the connection between neural areas such as V1 or the somatosensory cortex and vision or other forms of sense perception. Significant work has also been done on the importance of the interlaminar nucleus to the sleep/wake cycle and to dream states.

Brain-specific research of this sort has hugely increased our understanding of these processes, and the progress we are making toward mapping the neural underpinnings of mental phenomena like visual perception and sleep is extremely exciting. In recent years computer scientists have come much closer to building functional quantum-computation devices, which many hope will mimic brain behavior more accurately than do current digital computers. And psychologists have reached a better understanding of the differences between implicit and explicit mental processes such as memory and perception, as well as between automated and attended processes—differences, which many believe, will prove crucial to distinguishing conscious from unconscious activity.[3] So the core disciplines of neurobiology and cognitive psychology, as well as ancillary disciplines such as linguistics, computer science, organic chemistry, and even (if you subscribe to a particular minority-held view of the relation between mind and brain) quantum physics have been, and will certainly continue to be, indispensable for anyone who studies consciousness.[4] But all too often psychology and the hard sciences are treated as the only available paths to under-

[3]For a thorough and comprehensible discussion of these distinctions and the light they shed on unconscious processes, see Kihlstrom [52].

[4]For an elegant and much expanded argument along similar lines, see Flanagan's defense of what he dubs his "natural method" of studying consciousness, which involves triangulating between introspection, neuroscience, and cognitive psychology [33, ch. 1].

standing. As Crick notes with some glee, "everyone likes to show that philosophers are wrong," [21, p. 54] and I fear that this attitude has blinded many to the insights that a philosophical approach to consciousness can yield.

Some readers may be uncomfortable with the fact that many of my arguments and comments will rely rather heavily on introspection, or the examination of one's own mental states, attitudes, and experiences. Introspection as a tool for philosophical or psychological research is sometimes derided as unscientific, unreliable, and unhelpful. Although there are occasions where those claims are valid, I think it can be an extremely valuable tool if used carefully. Also, I can allow for at least one of these complaints—i.e., I can admit that introspection may not always be totally reliable—without damaging my later arguments. I realize that certain episodes of introspection, namely those that involve memory of perceptions rather than immediate perception, can occasionally produce varying results. Here is an example. If I am asked whether my car or my front door is closer to pure white in color, I would likely answer by turning my gaze inward and "looking" at memories of these two things. As I would be examining my own mental states, this seems as clear a case of introspection as I can imagine. Yet because my memory of these two colors is somewhat faint, and because it can be very difficult to compare qualia that are not immediately present, I might very well answer the question differently at different times. This case seems completely plausible, and so I can only conclude that introspection of memories is not incorrigible. For the record, I believe introspection of occurrent qualia and mental states are much more reliable, and in fact may be incorrigible. But this fallibility does not trouble me, for the scenario I have described does not threaten the existence of qualia. It only suggests

that we may have imperfect recall of them later on. So as long as we don't depend on the reliability of a particular form of introspection (namely, recall of qualitative memories), this corrigibility does not give us a reason to avoid the use of introspection as a tool for analysis. I do not have space to vindicate the general practice of introspection here, but I invite examination of exceptionally thorough and persuasive defenses of introspection by Charles Siewert [99, chs. 1, 2] and Leopold Stubenberg [104, ch. 3]. The problem of consciousness is hard enough that it would be foolish of us not to use every analytic instrument we can get our hands on and every strategy that seems even remotely plausible (though I hope my arguments prove to be a good deal more robust than this).

Not only is the problem of consciousness extraordinarily difficult, but it is also of paramount importance. Its importance stems from the number of philosophical and extra-philosophical issues and questions upon which it impinges. This is a trait of many problems in philosophy, but it seems especially prominent in the case of consciousness. For example, a full explanation of the nature of consciousness should shed light on the question of what sorts of biological or mechanical systems have or are capable of having consciousness, and this in turn affects issues in the areas of ethics, rights, and personhood. And a clear understanding of the function of consciousness will help speed work on free will and on issues in artificial intelligence.

The remainder of this dissertation will be broken into four main parts. In chapter two I will present a definition of consciousness, built around the notion of qualia. Chapter three will expand on this definition, filling in some gaps that remain from chapter two. Chapter four will address the issue of machine consciousness, and chapter five will examine

a few of the implications of the claims I have made in the previous chapters, focusing especially on personal identity and personal rights.

One further comment about the structure of this project is in order. Instead of doggedly pursuing a single line of thought or argument, I will present a collection of ideas and arguments which, although not tightly integrated, do support each other and contribute to a single overarching theory of consciousness. Specifically, they are all bound together by my general theory of a qualia-based conception of consciousness. Let us turn to that conception now.

Chapter 2

What is consciousness?

> *There's something queer about describing consciousness: whatever people mean to say, they just can't seem to make it clear. It's not like feeling confused or ignorant. Instead, we feel we know what's going on but can't describe it properly. How could anything seem so close, yet always keep beyond our reach?*
>
> Marvin Minsky [67, p. 151]

2.1 A qualia-based definition

"Consciousness" is indeed a slippery term. Philosophers, psychologists, computer scientists, and academics of other stripes each have their own definition of the word, and different shades of meaning are found even within individual disciplines. This problem is especially rampant in philosophy, where untold numbers of unhelpful articles and books have been written because of misunderstandings resulting from different uses of the term. In order to minimize such confusion in this dissertation, my first task will be to set forth as straightforward an explanation as possible of how I intend to use the word. I should make clear that I consider the word "consciousness" to be, in an important sense, a term

of art: we can declare that it means almost anything, and scholars in different fields do in fact use it in ways that emphasize what is especially important to their disciplines. Now I have already stated that I have a clear agenda in this dissertation: I want to think about what consciousness is and what, if anything, it does for us. But in recognizing the enormous variety of ways in which the term can be used, I seem to have made things very difficult for myself from the outset. After all, if anyone can define consciousness in any way they please, what is to prevent us from ending up with countless conflicting—but all equally correct—answers to these questions?

While it is true that "consciousness" can mean different things to different people, I think there is a central concept, or a core notion, that underlies the majority of the standard ways in which we use the term. It is this notion that I will do my best to reveal and discuss. Philosophers are well known both for doubting the existence of phenomena or things that were never the least bit suspect to those untrained in philosophy (think of the skepticism with which many philosophers view simple concepts like *inductive knowledge* or *folk psychology*—skepticism that no doubt strikes most people as utterly bizarre), and for taking well-known concepts that are easily and securely used in normal, casual conversation and using them in ways that render them virtually unrecognizable to laymen (consider certain philosophical analyses of apparently simple concepts such as *meaning*, *virtue*, or *scientific explanation*). Certainly there are those who have either denied the existence of or produced massively confused and twisted ideas of consciousness. I point this out not to legitimize these approaches, but rather to say that I do not know where even to begin to deal with them. As a result, I will simply assume these positions to be wrong. I think that

consciousness does exist, and that it exists in a form very similar to the standard notion most people either presently have or else would have if they gave the subject even minimal attention. This is not to say that it is easy to describe this notion; indeed, giving such a description is the very first task I have set for myself in this chapter. But I will not attempt either to defend it against those who deny its existence or to persuade those who use the term in ways completely divorced from the standard range of meanings that we take it to have. If you don't think that you are conscious, or if you don't realize that you are conscious in the fundamental way that the rest of us are, then I can offer only sympathy, and not arguments.

This core notion of consciousness (which I have not yet defined) is the component of the standard array of meanings that I take to be the most interesting and troubling. Perhaps more importantly, this notion allows us to get a new grip on a number of old philosophical problems. Consciousness is a fascinating topic for research even when it is considered in isolation of other issues, but its full importance cannot be understood without seeing how it relates to other philosophical questions, such as those involving personhood and personal rights. And the definition I provide will give us a wedge with which to pry into these topics. I will show in chapter three that other definitions either reduce to the definition I have given, or else do not afford us the same insight into these problems, if they offer any at all. So though some readers may object to my definition of consciousness, I hope they will agree that the phenomenon to which my use of the term refers is a particularly puzzling one, and that this phenomenon is especially worthy of further philosophical investigation due to its connections to other philosophical issues.

But enough preamble. My definition of consciousness can be stated quite simply: consciousness is the experiencing of *qualia*, a term I will define momentarily. A system must continue to experience qualia if it is to remain conscious; any periods during which no qualia are experienced are periods in which the system has lost consciousness. However, this slipping in and out of consciousness is not problematic. We do it every day when we fall into and out of dreamless (hence qualia-less) sleep. There are many ways to use the terms "conscious" or "consciousness" that conform to this definition.

- He came out of the coma and regained consciousness.
- I became conscious of a gnawing hunger that had come on surprisingly swiftly.
- She is self-conscious about her height.
- Anxiety about my upcoming recital eventually emerged from my subconscious, becoming vividly and horribly conscious.

At the moment I do not want to delve any further into how these uses of "conscious" may differ, but it is important to see that they all share at least one common element: they all involve qualia.

What do I mean by "qualia"? The term is thought to have originated with C. S. Peirce,[1] but only fairly recently has it gained wide currency among philosophers of mind. Discussions of qualia often begin with the claim, made in a weary and despairing tone, that it is impossible to describe what qualia are. Any efforts at description or definition are then

[1] "There is a distinctive *quale* to every combination of sensations—There is a distinctive *quale* to every work of art—a distinctive *quale* to this moment as it is to me—a peculiar *quale* to every day and every week—a peculiar *quale* to my whole personal consciousness. I appeal to your introspection to bear me out." [75, par. 223] (emphasis in the original)

traditionally abandoned in favor of ostensive gestures toward individual qualia. Witness two classic examples of this strategy, the first by Ned Block (using the term "phenomenal consciousness" essentially synonymously with "qualia"), and the second by David Chalmers:

> I cannot define phenomenal consciousness in any remotely non-circular way. I don't consider this an embarrassment. The history of reductive definitions in philosophy should lead one not to expect a reductive definition of anything. But the best one can do for phenomenal consciousness is in some respects worse than for many other things because really all one can do is point to the phenomenon. [9, p. 230]

> What is central to consciousness, at least in the most interesting sense, is experience. But this is not a definition. At best, it is clarification. Trying to define conscious experience in terms of more primitive notions is fruitless. One might as well try to define matter or space in terms of something more fundamental. The best we can do is give illustrations and characterizations that lie at the same level. [15, p. 4]

But I do not understand why it is considered so difficult to define the general nature of qualia. Providing a complete description of an individual *quale* (singular of "qualia") seems well nigh impossible, I agree. But I just want to show what sort of things I'm talking about when I raise the topic of qualia, and that task doesn't seem so onerous. Unfortunately I can only define qualia using a series of phrases that will be excruciatingly familiar to anyone versed in contemporary philosophy of mind, but so be it. A quale is the particular *way it seems* to a person to see something, taste something, or get sensory input from any of his three other sensory modalities. Qualia are the *raw feels* associated with experiences; they are what make different experiences seem or feel different from one another. A quale is *what it is like* to undergo an experience. A quale is what gives an experience its *subjective* element. Qualia are the *phenomenal contents* of experiences. Or to borrow a sublimely simple illustration of the concept from Stubenberg, "the fact that there is something it is like to be you consists in the fact that you have qualia." [104, p. 5] Note that so far I have

defined qualia purely in terms of sensory experience—all of the examples of qualia I have given are brought on by sense perception. However, I will eventually argue that all mental states that we would prephilosophically consider to be conscious (i.e., all of those states that are not deeply and permanently unconscious in some Freudian sense) involve qualia. For instance, I claim (and will later demonstrate) that common, everyday thoughts such as "that chair is orange" or "it's often windy in Boston" essentially involve qualia just as sense perception does. But for the sake of simplicity I now want to discuss only this thinner, perhaps less controversial notion of qualia.

Definitions are always clearer when supplemented with examples, so I will now point to some common, everyday qualia. The qualia associated with eating an apple are what let you know that you are eating an apple and not a pear or a taco. Apple-eating involves a particular class of tastes, a particular class of smells, a particular class of tactile textures both in the hand and in the mouth, and a class of particular sounds. If you look at the apple as you eat it, there will also be a class of particular visual experiences involved. I am careful to specify that classes rather than individuals are involved because while presumably no two apples taste exactly alike, their flavor will be very similar in certain important respects. The set of tastes produced by all apples will then compose the class of apple-tastes. Pear-eating will involve similar tastes, smells, etc., some of which may be quite close to the experiences involved with apple-eating. But there will be enough differences to allow anyone to tell whether they are eating an apple or a pear. The qualia that accompany taco-eating will generally be even farther removed. These characteristic experiences—both those that are similar and those that are different—are qualia. Another example: qualia are

what distinguish a person's perception of a green disk of light projected on a wall from her perception of an otherwise identical red disk of light. At least, qualia are what an average observer uses to distinguish the two. Science tells us that each disk is made up of light of a different wavelength reflecting off the wall's surface, and our experience tells us that the disks differ in certain other properties (e.g., the latter is capable of making a bull charge or a driver stop, while the former does not). But the most immediate and striking difference between them is that *they look to be of different colors*, and this difference is nothing more or less than the difference in the qualia they produce in us.

Qualia vary enormously even within a single sensory modality. In the tactile realm, separate qualia are associated with sharp pains, dull aches, tickles, prickles, itches, aches, stickiness, coolness, warmth, vibration of various frequencies, pressure, and all of the other sensations that come from agitation of the skin. A normal eye can distinguish millions of different colors, and taste buds can distinguish nearly as many combinations of salty, sweet, sour, and bitter flavors. Different people have varying capacities for discriminating among sounds, but experienced musicians' capacities for differentiating between similar pitches and timbres are extraordinary. The sense of smell is probably the least discriminating of the sensory modalities (at least among humans), but even the lowly nose has the capacity to differentiate among a bewildering array of smells. The range of qualia that we can enjoy—or suffer from—seems virtually limitless.

I will end this definition of qualia with a crucial metaphysical point. I have said in previous paragraphs that qualia are experiences, and I think this is important to emphasize. Specifically, it is important to contrast this with the false notion that one experiences

qualia. While it is true that we experience events or things, I claim that these experiences are themselves made up of qualia. Experiences *comprise* qualia; they are not *of* qualia. However, for ease of discussion, I will sometimes use the technically incorrect locution "to experience a quale."

We can see that this locution is, strictly speaking, inaccurate once we get clear on the ontological status of qualia. But this is a difficult issue. We might first ask if they are objects or properties. If they are objects, are they objects in the mind (however one might decide to cash out that concept) or objects in the world outside the mind? If they are properties, are they properties of objects or properties of mental states?[2] Arguments on this issue have been bandied about in the literature for years, and I will not try to settle the matter here. In fact, the exact ontology of qualia is largely irrelevant to the claims I will make about them. Nevertheless, I cast my vote for, and through the course of this dissertation will assume, the last of these four option. Qualia are properties of mental states, and as such exist entirely within the mind. I believe their privacy and ontological subjectivity (properties that I will discuss at length below) rule out the possibility that they are material objects (i.e., objects that can exist independently of experience), and the fact that the same object can produce vastly different qualia in different people, or in the same person at different times, or even in the same person at the same time, suggests that they are not properties of any external object.[3] Examples of the three cases are as follows.

[2]A third option is sometimes suggested. William Lycan [60, p. 84], for example, discusses—but ultimately rejects—the possibility that qualia are properties not of objects or mental states, but of phenomenal particulars. Not knowing how to make sense of the concept of a phenomenal particular, I will ignore this position.

[3]For support of the view that qualia are properties of external objects and not of mental states, see Harman [44, 45]. Loar [57], Lycan [60, p. 88], Tye [109] and Dretske [30] hold positions similar to that of Harman. Block [8] gives a comprehensive rebuttal of Harman using an argument somewhat different from mine.

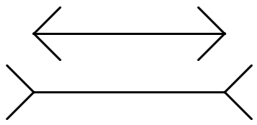

Figure 2.1: Müller-Lyer illusion

First, Jones, who has just come inside after spending time outdoors on a brilliantly lit day, might look at a ripe plum and see it as dull gray, whereas Smith, emerging at the same time from a darkened movie theater, might see the same plum under the same lighting conditions as bright purple. Second, if Jones were to eat the plum after eating something salty, sweetness might be the most vivid quale it produces in him, whereas if he were to take another bite of plum later, after eating ice cream, the most prominent quale resulting from the plum might be that of sourness. For an example of the third case, think of the Müller-Lyer illusion, in which two lines of identical length produce different qualia in a single observer due to differently angled "fins" attached to the lines (see Figure 2.1). This leaves only the possibility that qualia are properties of mental states. Specifically, I want to claim that qualia are tropes, where I use "tropes" in the traditional sense of being instances of property universals.[4]

At this point I should interrupt my talk of qualia in order to say a few words about mental states. My claim that qualia are properties of mental states does us little good unless we have a thorough understanding of what a mental state is. Here I regret to

[4]The metaphysical use of the term "trope" was introduced by Donald C. Williams in 1953 [118]. Other examples include Arnold Schwarzenegger's confidence (a particular instance of the universal property of confidence) and my Saab's whiteness (an instance of some universal whiteness). Tropes are somewhat controversial metaphysical entities, and although I think they can be used fruitfully in an analysis of consciousness, I wish to remain agnostic on the exact ontological status of tropes, universals, and particulars. For a variety of views on this issue, see Armstrong [3], Leibniz [55, §§9, 14], Locke [58, p. 159], and Stout [103].

say that I can give only a circular definition: a mental state is an occurrent element of one's mind at a particular time. For example, if I am hungry at noon on my 30th birthday, then I have the mental state of hunger at that time. If ten minutes later I find myself imagining the feeling of a warm breeze and the sound of rustling palm fronds on a Tahitian beach, then I am in a mental state of imagining that I am in Tahiti. Note that this requires that I realize that it is mere imagination—if those sensations were real, and were caused by my actually being on a Tahitian beach, then I would be in the mental state of enjoying being on a Tahitian beach. And if I was in fact sitting at my desk in Berkeley while thinking that those sensations were real (the result of an extraordinarily vivid and powerful daydream, say), then I'm honestly not sure how to characterize the resulting mental state.

Although I have given a somewhat vague definition, the definition conforms to common sense and should be readily comprehensible. And I believe it offers all the precision we need. But a related question remains: how should we individuate mental states? At any given time there may be several elements present in my mind. As I write this I feel uncomfortably warm, I feel an itch on the bridge of my nose, and I feel mildly anxious about a looming deadline. Am I in one mental state or three? And what happens if my anxiety grows into full-fledged panic—have I then shifted into an entirely new mental state, or has my original mental state merely strengthened in some sense? The question of individuation is enormously difficult, but I think it is not a question that need be answered in order for consciousness to be discussed fruitfully. The question reveals the vague boundaries around the notion of a mental state, but I see no reason why the nebulous shape and limits of individual mental states should pose any real problem, for qualia can always be thought

of as properties of the overall mental state one is in. Combine all of the elements of a person's mind at a particular time, and if any of those elements have qualitative properties then we can say that that person is experiencing qualia. It doesn't matter what the exact components of the overall mental state are. If I am hungry, then the quale of hunger is a property of my overall mental state, regardless of what (if any) other thoughts, perceptions, moods, emotions, etc. are flitting through my mind at the same time. The hunger-quale is a property of the narrow and specific mental state of hunger, but it is also a property of a general mental state of misery, which might comprise elements of hunger, exhaustion, anxiety, and a pounding headache.

Let us now leave the issue of individuation and return to the ontological status of qualia. I have claimed that each token quale is a property, and is an instance of a general property that is a qualia-type. This means that two people can have qualitatively identical (though not numerically identical, as will be discussed below) qualia if each quale is a token of the same general qualia-type. This classification of qualia as properties, and specifically as tropes, helps us understand certain properties of qualia themselves—properties that are hard to account for on other classifications. I will now examine these properties in more depth.

The general notion of qualia is easy to define, but there are three important properties of qualia that make it difficult to identify and talk about any particular quale: they are *private*, they are *ontologically subjective*, and they are *ineffable*. Let's look first at privacy. A token quale is private in the sense that that particular quale can be possessed only by one person. The pain you feel when stubbing your toe is yours alone—there is

no scientific apparatus, no common or esoteric mental technique, no physical act that will enable anyone else to feel that particular pain.[5] By stubbing my own toe I may be able to experience a pain that is quite similar to your pain, and may even be type-identical to yours (i.e., it may feel to me exactly the way your pain feels to you), but I cannot feel *your token pain*. Your mental states are securely locked within your own mind, forever inaccessible to other minds. Qualia are not like buildings, cars, or other physical objects, in the sense that they are not universally accessible like those other objects are. Anyone can see the Tower of London simply by standing in the right place and opening his or her eyes, but an individual quale that is a property of a mental state of a particular mind resists even the most valiant efforts of other minds to experience it. It is important to emphasize that I am not claiming that two people cannot have similar or even qualitatively identical qualia. To make such a claim one would have to maintain not only that my pains feel nothing at all like yours, but also that my pains feel nothing at all like *any* of your qualia—a claim that could conceivably be true, but which virtually all biological, psychological, and behavioral evidence argues against. I am simply saying that two people cannot experience the same particular quale, or the same trope. It also seems unlikely that a single person could experience two token-identical qualia as long as they are temporally separated.

I believe the first interesting property of qualia—their privacy—derives from their second interesting property—their *ontological subjectivity*. This property was first discussed

[5]Privacy seems to me the most apparent and indisputable property of qualia, and the literature is replete with arguments to this effect (almost certainly the best known of these is by Thomas Nagel [69]). But as is typical in philosophy, there are dissenters of even what seems to be an obviously true claim. Flanagan [33, ch. 5], for one, gives a long but less than transparent argument that what it is like to be me is not closed off to you, or as he puts it, "the subjectivity of qualia is not impenetrable" or later, "particular subjects of experience can be known as such by other subjects of experience." This is an intriguing view to be sure, but I'm not sure I understand his argument well enough to feel comfortable discussing or assessing it.

at length (as far as I know) by John Searle.[6] Things with this property depend on minds for their existence. Things that exist independently of minds, in contrast, are *ontologically objective*. Qualia are ontologically subjective objects par excellence (other examples might include the score of a baseball game or the resale value of my old Saab), while buildings and wristwatches are ontologically objective. This is not to say that the qualia produced in me when I eat strawberries are any less real than the Chrysler Building or my watch. Although these things have different ontological status, they all exist in the world and are all equally real. It is important to emphasize this parity, for the term "subjective" is sometimes taken to indicate a certain lack of reality. When we categorize opinions about music as "subjective" we mean that they are not ultimately true or false. But it is important to distinguish between—and here I borrow more of Searle's terminology—things that are *epistemically* subjective and things that are *ontologically* subjective. Musical opinions are an example of the former. Most of us (though probably not all, admittedly) would claim that there is no truth of the matter about whether Prokofiev's First Symphony is a good piece of music, so any aesthetic claims made about it are subjective in some sense. It is this sense that Searle dubs "epistemic subjectivity." I believe that it is a wonderful symphony, but somebody else might believe, just as legitimately, that it is trite and childish. Ontological subjectivity is very different. Ontologically objective things enjoy no advantage of security or truth over ontologically subjective things. The latter are just as real as the former, but happen to belong to a different ontological category. The experience of orange I come to have when looking at a carrot is ontologically subjective, whereas the carrot itself is

[6] See Searle [96] for a much more thorough treatment of ontological subjectivity and its cousins (discussed below): ontological objectivity, epistemic objectivity, and epistemic subjectivity.

ontologically objective. There are facts about both things, both things can be studied, and both things have causal powers. Both are equally real, but they are fundamentally different *sorts* of things. One is a physical object that exists in the real world; the other is a mental object which, although it also exists in the real world, exists only within a special part of the world—namely, within a human mind. The carrot could exist if no minds did, but the orangeness (and the carrot's taste, and the sound it makes when it is bit into, and the tactile feel it produces on the tongue) could not. My point in raising the issue of ontological subjectivity is this: I think that ontological subjectivity is a necessary and probably sufficient condition on privacy.

Let's look first at the *necessary* part of this claim. I am assuming that all things are either ontologically objective or ontologically subjective. Then anything that is not ontologically subjective must exist freely in the world, independent of any mind. This suggests that it is publicly accessible to any mind with the appropriate sense organs. That is, it is not private in the sense in which I have used the term to describe qualia. Just as everyone has perceptual access to a particular car or building, anyone would have perceptual access to a particular *ontologically objective* quale (though I don't for a minute think such a thing could exist). Therefore ontological subjectivity seems to be a necessary condition for privacy.

It is very likely also a *sufficient* condition. I am inclined to think it is sufficient simply because it is hard to imagine how minds other than the mind on which the ontologically subjective thing depends could have access to that thing. Short of extra-sensory perception or other decidedly unscientific methods of communication, there don't appear

to be any means by which an ontologically subjective thing could impinge directly on other minds. I could perhaps access an ontologically subjective thing that is a reasonable facsimile of an ontologically subjective thing of yours, but the thing that I am accessing is *mine*, and is distinct from yours. We can see that this is so from the following scenario. Should one brain be destroyed, all of the ontologically subjective things associated with that brain would presumably be destroyed as well. But in such a case there is no reason to think that the other, non-destroyed brain would also lose its corresponding ontologically subjective object. You could be feeling an ontologically subjective pain at the same time that I feel a type-identical pain, and if my brain were destroyed, you would go right on feeling that token pain. So I argue that one would be mistaken to think that two people could have direct access to a single ontologically subjective thing. Cases where this appears to be the case should instead be described as cases in which there are two distinct (albeit type-identical) ontologically subjective things, and each person has direct access to only one of these things. This concludes my argument that ontological subjectivity is probably sufficient for privacy.

Why am I not completely convinced by this argument—why do I insist that ontological subjectivity is merely *probably* sufficient for privacy? The reason is that there are possible counterexamples that make me reluctant to endorse the claim fully. These counterexamples are not terribly convincing, but they are enough to give one pause. They revolve around the uncertain ontological status of moral truths and numbers: some moral subjectivists claim that moral truths require minds for existence, and mathematical intuitionalists make the same claim about numbers. Moral truths and numbers could then

perhaps be thought of as ontologically subjective (for according to these views, they are as mind-dependent as, say, pain) but freely accessible to all, and so not private. That is, it could be argued that you have exactly the same concept of the number three as I do, or perhaps the same sense that theft is morally wrong, even though these things are essentially ontologically subjective. The argument would also seem to require that a single token be involved in each case: my concept of three or moral sense that theft is wrong would have to be *token-identical* to yours, whatever that might mean. If this is true, then ontological subjectivity seems not to be sufficient to ensure privacy. But because the metaphysics involved in mathematics and moral philosophy are highly controversial and because this sort of argument would require justification of several dubious steps, I don't want to give these putative counterexamples too much weight. I still maintain that the most sensible route to take is to insist that the ontological subjectivity of qualia is necessary, and very probably sufficient, for their privacy.

One might wonder why I treat ontological subjectivity and privacy differently. After all, if x is a necessary and sufficient condition for y, then aren't x and y the same thing? The answer to this is that even if the labels x and y are coreferential, these labels can highlight different aspects of that thing. "Morning star" emphasizes that the star to which it refers is visible in the morning, whereas the coreferring term "evening star" indicates that (what happens to be) the same star is visible in the evening. Similarly, the terms "private" and "ontologically subjective" focus on slightly different aspects of the same basic trait. The former stresses the mind-dependence of a thing, whereas the latter draws attention to the inaccessibility of that thing to other minds. It is not immediately obvious that all

ontologically subjective things must be private, nor is it readily apparent that all private things must be ontologically subjective. So while I have argued that these two traits do in fact go together, which suggests that they are ultimately two descriptions of a single property, I do think it is useful to pull these terms apart from each other and treat them separately.

But let us return to my argument that qualia are private. There would seem to be plenty of room to dispute this claim. For might not a blank piece of paper produce exactly the same quale of whiteness in you as it does in me? Couldn't a chirping bird produce identical sonic qualia in anyone within earshot? One of the best known and most interesting arguments in this direction is given by Daniel Dennett [24]. The role of his argument is a bit strange in that its aim is not so much to persuade us of the non-private nature of qualia as it is to prepare us for Dennett's ultimate thesis that the experiencing of qualia (hence, the phenomenon of consciousness itself) is illusory—it is really nothing more than a useful "stance" that we take toward certain human and non-human systems. But the initial argument he makes is a fairly common one, and thus is worth considering independently of his larger (and by my lights, thoroughly misguided) view of the mind. Now, it seems to me there are two ways to deal responses like Dennett's. First, one could deny that the qualia had by two people can be qualitatively identical, much less numerically identical. This approach comports with the common sense view that no two people perceive the same thing in exactly the same way. Perhaps the cones in your retinae are more sensitive than mine, so your eyes pick up a slight yellow tinge to the paper that I miss. Exposure to loud music has damaged the cilia in my ears, so I miss high-frequency elements of the birdsong

that may come through loud and clear for you. But my opponent might then suggest that we could imagine a case in which two people with identically functioning sense organs see the paper in exactly the same light, at exactly the same angle, etc. Or imagine two people who are standing equidistant from a single bird, with no wind to produce asymmetrical acoustic effects, etc. Couldn't we then say that the two subjects really do experience exactly the same qualia? Perhaps, but then we can invoke the second response to this argument: while two people can share qualitatively identical qualia, they cannot share numerically identical qualia. The objection to my view does not demonstrate that different people have access to the same token quale, since it could be that they just have access to different qualia tokens that are a lot alike. Furthermore, the scenario I presented above in which a quale continues in one mind after another mind that has a similar quale is destroyed or drops into a coma suggests that it is far more likely that the two qualia are merely qualitatively—not numerically—identical. It seems trivially true that numerically identical objects must share all of their properties,[7] and there is no reason to doubt this principle would hold for numerically identical properties as well as numerically identical objects. Since the properties of your qualia can change while mine remain constant (or to return to the most extreme case, yours may disappear entirely while I continue to experience mine), this identity principle makes it clear that our similar-seeming qualia can only be qualitatively identical, and not numerically identical. In other words, my bird-song quale and your bird-song quale are distinct tokens of a particular quale-type rather than being a single, shared quale-token.

I have claimed several times now that the possibility of a quale vanishing from one

[7]This principle is sometimes considered to be a form of Leibniz' Law [56], which states that two things that share all properties must be numerically identical. However, the two principles are quite different and should not be confused: the former asserts the indiscernability of identicals whereas the latter asserts the identity of indiscernables.

mind while a similar quale continues in another mind suggests that different people do not experience token-identical qualia. But one might respond by saying that such a situation would be better described by saying that two people have access to the same quale of whiteness, and when one person becomes comatose, dies, or for some other reason ceases to experience that quale, that person simply *loses access* to that quale. This description would allow the quale's properties to stay constant (though its relations to objects would change), which in turn allows us to say that there is only one quale involved. This suggests that qualia are not private. But this characterization of qualia is inconsistent with our description of them as properties of mental states. If a given quale is a property of a particular mental state, and if that mental state is a contained in or realized in a particular brain, then it is nonsensical to say that two people share access to a single quale. Which mental state is that quale a property of, and which brain is that mental state realized in? We can't say that the quale belongs simultaneously to two minds, but if we pick just one mind to "house" the quale, we run into a problem of asymmetry—we have no grounds for picking that mind over any other. Thus, the privacy of qualia seems to fall directly out of their metaphysical status as properties of mental states.

Let us now turn away from privacy and ontological subjectivity, and toward the third peculiar property of qualia, their ineffability. By this I mean that qualia are not precisely or fully describable in language. Further, unless one resorts to comparing a particular quale to other qualia (e.g., orange things look like a cross between red things and yellow things, or the texture of a raw potato on the tongue is similar to that of a raw apple), they seem not to be describable at all. Imagine trying to convey the experience of sight to a

person who has been blind since birth; one is baffled by where even to begin. It is no easier to describe qualia produced by any of the other sensory modalities.

Qualia are not unique to perception: as I will argue in chapter three, propositional thoughts—thoughts such as "Today is Monday" or "I want spaghetti for dinner"—involve qualia that are just as vivid as those involved with sense impressions.[8] They also resist description just as strongly as do qualia involved with sense impressions. Prima facie, it may seem easy to describe the qualia associated with my thought that today is Monday or my desire for spaghetti. I simply ask you to imagine that today is Monday, or to imagine that you want spaghetti. This technique could perhaps allow you to summon up qualia that are type-identical to my Monday-believing quale and spaghetti-craving quale. But this is not a credible counterexample to the claim that qualia are ineffable, for I have described the qualia in a roundabout way rather than describing them directly. Rather than describing the qualia themselves, I have conveyed to you the semantic component of my mental states. You were then able to conjure up mental states with similar semantic content, which in turn produced qualia in you. So although you may end up with qualia that are type-identical to mine, the qualia themselves have not been described. Rather, I have described a way of producing in yourself qualia similar to mine.

Another way to see that qualia of both sorts—those produced by sensory perception and those that accompany propositional thought—are ineffable is to consider that qualia exist in degrees, and it is not clear how one could specify the strength of a particular quale. How do I express exactly how much I want spaghetti for dinner, let alone how much

[8] At least, this claim holds for propositional thoughts that we are aware of; unconscious propositional thoughts are exempt. More on this in chapter three.

better spaghetti sounds than sushi? How might I convey how certain I am that today is Monday? And even if we somehow did develop techniques for expressing these sorts of things with arbitrary precision, it is hard to see how we could test the similarity between my qualia in me and the qualia my description produces in you. If qualia were not ineffable, then a trivial test could be performed. I could simply ask you to describe your qualia, and then compare that description to a description of my own qualia.

There is another fact that contributes to our difficulty in describing the qualia that accompany propositional thoughts in particular: propositional thoughts are holistic. Note that this is not a claim about a logical property of qualia, but rather a claim about the settings in which they often occur. Also, I am not going to argue that this fact is sufficient to make qualia fully ineffable; my claim is simply that it contributes to the difficulty we have in describing qualia. So this is weaker than the previous arguments I have made about ineffability in these two important ways. Now what is my argument exactly? In saying that propositional thoughts are holistic, I mean that most thoughts are tightly interconnected to other thoughts, and the strength and number of such connections can make it difficult to identify individual thoughts. The difficulty or impossibility of individuating thoughts translates into a difficulty separating the qualia that accompany one thought from the qualia that accompany others. This in itself does not obviously contribute to our difficulty in describing qualia, but the fact that two people can have very different networks of belief or thought connected to a single common thought does. Even if I were able to identify and convey to you what my core thought is at a given time (e.g., I am thinking about Mondays in general), the holistic nature of propositional thought virtually ensures that the network

of thoughts produced in you will vary significantly from the network produced in me. The qualia that accompany your network will, in turn, differ from the qualia that accompany my network. My thinking-of-Monday qualia might include anxiety and dread, while yours might include anticipation and excitement. Again, this does not mean that the qualia that accompany my Monday-thought are, strictly speaking, ineffable, but it does suggest that I will have difficulty conveying exactly which qualia are associated with my Monday-thought. I cannot simply say that my qualia are similar to those produced in you when you think about Mondays, for your Monday-thought may be connected to very different thoughts than mine is. So I would instead have to trace my holistic network out from my Monday-thought, describing each quale encountered along the way. This would be a difficult (and perhaps even infinite) task.

The problem introduced by the holistic nature of propositional thought applies not just to two people at the same or different times, but also to a single person at different times. The constant flux of qualia produced even by a single thought over time is one reason why qualia are often hard to describe even to ourselves. When we are faced only with a fleeting series of variegated qualia, it becomes very difficult to pin down a single quale long enough to describe it. William James provides an excellent illustration of this holism:

> We feel things differently accordingly as we are sleepy or awake, hungry or full, fresh or tired; differently at night and in the morning, differently in summer and in winter; and above all, differently in childhood, manhood, and old age. And yet we never doubt that our feelings reveal the same world, with the same sensible qualities and the same sensible things occupying it. [50, p. 22]

Though his emphasis here is on the holism of sense perception rather than thought, his comments apply equally well to propositional thought (indeed, he later notes that we can

never have the same idea twice). There will be slightly different networks of related thought and qualia associated with different instances of the same person thinking "I'm hungry" or "Today is Monday."[9]

There is perhaps even some reason to believe that *all* qualia, and not just qualia associated with propositional thought, are similarly holistic. Or to refine the claim, it could be that consciousness can only be understood by looking at the ways in which qualia interact and interconnect; looking at a single quale in isolation will not help unravel the mysteries of consciousness. Robert van Gulick has this idea in mind when he urges us to use what he calls a Kantian notion of qualia. According to Van Gulick, the term "qualia" when used in this way doesn't refer just to a succession of raw feels, but rather to "the organized cognitive experience of a world of objects and of ourselves as subjects within that world." Focusing just on raw feels would be a mistake:

> First, it would provide too narrow a definition of what needs to be explained, and, secondly, I doubt that qualia and raw feels can themselves be understood in isolation from the roles they play within the richer Kantian structure of phenomenal experience. [111, p. 137]

I am inclined to disagree with this view primarily on the grounds that merely understanding the nature of qualia is a problem of broad enough scope already, and that we have no need

[9]One might be tempted to say that this point accords with the almost universally accepted view that the mind is supervenient in some sense on the brain. That is, we might want to claim that some relevant part of the brain would have to be in an identical neurological state on each occasion of a particular propositional thought if those thoughts are to have exactly the same network of connected thoughts and involve exactly the same qualia, but given the staggering complexity of the brain and the rapid pace of change within it, it is extremely unlikely that sufficiently large portions of the brain would ever be in identical states. However, making such a claim would be to confuse the sufficient conditions for having a certain mental state with the necessary conditions for having that state. While a certain pattern of neural firing p may be sufficient to produce mental state m in a particular person, there is no reason to think that mental state m can *only* be produced by firing pattern p in that person; other patterns may be able to produce exactly the same mental state. This is even more apparent when we imagine mental state m occurring in other minds: different people often share qualitatively identical mental states but probably very rarely (if ever) have exactly the same neuronal firing patterns. Many proposed solutions to the mind/body problem draw explicitly on this "multiple realizability" of mental states.

yet to widen our investigation into the problem of how qualia contribute to one's experience as a subject within the world. But I make note of Van Gulick's position here to illustrate the holistic nature of qualia (albeit perhaps at a higher level of abstraction than the level I am investigating).

Even if I have shown that qualia are difficult to describe, this is not the same as showing that they cannot be described—that they are truly ineffable. And I admit that I probably cannot show that. I have no unassailable argument for the ineffability of qualia, but I do have quite a bit of experience trying unsuccessfully to convey directly the phenomenal features of qualia. I just don't see how it can be done. As alluded to earlier, I think the closest we can come is to describe a particular quale by locating it relative to other qualia. The dull beige color of my computer lies somewhere between the creamy yellow of my stationery and the cool gray of my carpet. Such descriptions are suspect from the outset because they rely on prior experience of certain qualia; your experience of my carpet may be quite different from mine. But these descriptions also face the problem of precision: where exactly on the spectrum between my stationery and my carpet does the color of my computer lie? Descriptions of this sort can eliminate huge chunks of qualia-space from consideration, but it is doubtful that they can home in on any one particular quale with the necessary precision. Again, this claim is supported not by an argument but rather by empirical evidence. Try to describe the qualia produced in you as you eat lunch or listen to music. Even if you are allowed to refer to other qualia in your description, you will probably find that you have remarkably little success at conveying the distinctive features of those qualia.

One might wonder why I am emphasizing the ineffability of qualia when all sorts of other things seem equally ineffable.[10] For instance, calling a table "red" hardly identifies its color uniquely, for there are countless shades included under general term "red." Even the claim that a table is made of wood is susceptible to this problem, for that leaves many aspects of its composition unspecified. What sort of wood is it made of? Is it completely wooden, or are there elements of metal or glue included as well? In calling a table red and wooden am I not just assimilating it into general categories rather than really getting at its essence? In other words, one could argue that the ineffability of qualia is not at all an unusual trait, and that it stems from a limitation of the language we use to describe things rather than any mysterious property intrinsic to qualia.

I admit that this argument does bring to light the futility of trying to describe anything in such a way that its essence is fully captured, and I think this does indeed reveal an unfortunate feature of language. But I think there is still a distinction we can draw between qualia and tables. The difference is that in describing tables, we are trying to get at the physical properties of an object, and to describe them in as much detail as is relevant for our purposes. So while I can never completely capture the physical structure and composition of the table with language, I can describe it with essentially arbitrary precision: I could ratchet up the level of detail in my description by mentioning that the table is made of Oregon Douglas fir, was finished with a pint of linseed oil, and contains six brass screws in addition to half an ounce of Elmer's professional wood glue, or whatever it takes to make the description appropriate given my reason for describing the table in the first place. In the case of qualia, I'm not sure this can be done. Saying that the pain caused

[10] I am grateful to John Searle for bringing this issue to my attention.

by the glass sliver in my foot falls somewhere between the pain of a burn and the ache of a stubbed toe helps narrow down the range of ways that the pain might feel, but that is about as far as we can go. In describing what hunger feels like, I'm not sure I can say much more than that it feels like what you feel when you are hungry. The reason for this difference is not surprising. The things that are capable of being described with arbitrary precision are all things that are ontologically objective. They are material objects that exist in the real world, and they have real physical properties that can be described with as much or as little specificity as desired (in many cases we can even achieve perfect precision relative to our aims: describing the Petronas Towers in Kuala Lumpur as 1476 feet high is as precise and unambiguous a description—of their height, at least—as anyone could wish for). But the ontological subjectivity of qualia prevent them from being described this way. Ordinary language is much better suited for describing physical objects; it just doesn't seem capable of getting much purchase on qualia. Why this should be exactly, I don't know. Perhaps language could be developed with an eye to accurate description of qualia, but it is hard to see how that project could make any progress other than as a purely private language, given that there is no way for an inventor of qualia-specific terms to demonstrate their meaning to anyone other than himself. In any case, the upshot of all of this is that ineffability, while not unique to qualia, is certainly a characteristic feature that sets them apart from most physically grounded things, facts, or phenomena.

I mentioned earlier that the privacy and ineffability of qualia make them difficult to study or discuss. It should be clear by now how these two properties (along with ontological subjectivity) conspire to make this so. The privacy of your qualia prevents me from accessing

them directly, and their ineffability generally prevents you from describing them in a precise enough way to be useful. You can perhaps give me a rough idea of your qualia—you might describe how Mozart's Jupiter Symphony sounds to you by calling it a combination of the grand themes of Beethoven and the intricate rhythms of Bach—but you can't describe those qualia with arbitrary precision or assess the accuracy of the qualia subsequently produced in those hearing the description.

This ends my extended definition of qualia. I have introduced qualia as a means of getting a better grip on what consciousness might be, so let us now examine various aspects of consciousness and their relation to qualia.

2.2 State vs. creature consciousness

I have given a definition of qualia, pointed to a few qualia as examples, and discussed some of their characteristic properties. I have also stipulated that consciousness is the experiencing of qualia. But as mentioned earlier, there are several ways of using the terms "conscious" or "consciousness" that are consonant with this definition. It is important to make the differences between these uses explicit. Failure to explain or even to understand which aspect of consciousness one is speaking of has led to enormous confusion in the philosophical and psychological literature. One of the most common confusions attending use of the term involves the distinction between what is sometimes referred to as "state consciousness" and what is often called "creature consciousness." It is important to get clear on what, if anything, this distinction amounts to. If there is a distinction, it would also be important to know which aspect one is speaking of when making claims about

consciousness. Although this is somewhat familiar terrain to many philosophers of mind, there is still no broad consensus about how to use these terms. Nevertheless, the taxonomy of aspects of consciousness that I am about to provide is probably as close as we can get (and that's not very close) to being the received view on this matter.[11]

At the outset, I should say that while I think the distinction between state and creature consciousness can be helpful in certain instances, it should not be taken too seriously. It is very important to realize that ultimately there is only one kind of consciousness, and that is the phenomenon that happens when your mental states have qualitative properties. State and creature consciousness should properly be thought of as aspects of this single, unified phenomenon. I cannot stress this point enough. Sometimes we speak of whole systems being conscious, while at other times we talk about individual mental states that are conscious, and it is worth going into more detail about what exactly we mean when we talk in these ways. We just need to remember that both of these uses of "consciousness" are, in the end, really concerned with qualia.

It is also important to stave off possible confusion by making clear that I do not mean this distinction to be the same as the distinction that is sometimes made between "representational consciousness" and "sensational consciousness."[12] While it is certainly appropriate to note that sensation and representation are very different phenomena that may well be accompanied by different sorts of qualitative feels, that is not the distinction I am trying to describe here. Nor am I referring to the purported (and I think massively mis-

[11]For similar—though not identical—taxonomies, see Dretske [29], Goldman [37], Güzeldere [40], Lycan [60], and Rosenthal [85]. Much of my terminology is borrowed from Dretske. Earlier discussions of largely the same topic take place in James [51] and Brentano [10].

[12]For an interesting analysis of this distinction, see Peacocke [74, pp. 4–26]. He argues that perception and sensation are more closely linked than we might at first think.

leading) distinction between "access consciousness" and "phenomenal consciousness."[13] It is often said that there are two kinds of consciousness: access consciousness (also sometimes called psychological consciousness), which is the aspect of the mind that drives behavior, and phenomenal consciousness, which the part of the mind that is involved with sense perception, emotions, and any other activity that has a phenomenal, what-it's-like component. Although I agree that it can be helpful to think of the mind as having these two components, I think it is just wrong to think of them both as forms of consciousness. The first seems to fit squarely in the realm of psychology and plays an important role in that domain. But I certainly see no reason to consider it to be a form of consciousness. So while the distinction is a good one, the names generally used to describe the two things being distinguished are terribly misleading.

The origin of the access/phenomenal consciousness distinction is easy to guess: early consciousness scholars were probably struck by the fact that many mental states have both phenomenal and psychological components. Take pain: it has an unmistakable feel (it hurts!) and an equally clear cognitive or psychological element (it drives me to seek relief, it causes me to avoid performing the pain-causing behavior again, and so forth). But the mistake then made was to think of these both as flavors of consciousness when in fact the latter is an entirely separate phenomenon. Both are of course equally worthy of study, but I will focus my efforts here strictly on phenomenal consciousness, leaving the unfortunately named "access consciousness" to psychologists and perhaps artificial intelligence researchers.

I should also comment on my use of the term "person" throughout this dissertation.

[13]This distinction is generally attributed to Block [9], although essentially the same distinction is pointed out in Jackendoff [49], Chalmers [15], and Burge [13].

Although I will ultimately argue that the range of systems that can be considered persons is quite limited, in the initial stages of this inquiry into consciousness I want to at least entertain the possibility that non-human systems can be considered persons. This means that until chapter five, in which I will present a full discussion of personhood and what ought to be classified as a person (and here is a preview of my unsurprising claim: you're a person in at least some minimal sense if you're conscious), I will be using the term in a broader sense than is typical. That is, unless specified otherwise, "person" should be thought of as *potentially* referring a wide range of physical systems, pending evidence of their consciousness. This includes human beings, all other animals (regardless of position on the phylogenetic scale), plants of all sorts, computers of virtually any architecture, and any other system that one has reason to believe might be conscious. Again, I will narrow this list considerably in chapter five.

2.2.1 Creature consciousness

Creature consciousness is a property of whole organisms. The term is often used to indicate that a person is awake, at least somewhat alert, and at least minimally aware of his surroundings. But I think this definition fails to get at what we are interested in when we speak of consciousness. I propose an alternative definition that I think will be more helpful: something is creature-conscious if and only if it is actively experiencing qualia. This definition has at least five advantages over the standard definition. First, it is a simpler, non-conjunctive definition. We generally have a simple concept in mind when speaking of creature consciousness, and a simple definition of the term seems to capture this concept more accurately than a sprawling, heterogeneous definition. Second, it is less

ambiguous. There may be puzzle cases where there seems to be no fact of the matter, even when considered from the first-person perspective, about whether a person is awake and alert. But there is rarely any doubt—especially from the first-person perspective—about whether a person is experiencing qualia. Third, my definition excludes systems that would qualify as creature-conscious under the standard definition but which we are inclined to consider non-conscious. For example, it is sometimes claimed that a standard personal computer equipped with perfectly ordinary peripherals such as a scanner and a microphone can be seen as awake, alert, and aware of its surroundings, but I know of almost no one who considers their computer to be conscious. Fourth, my definition includes systems that might not qualify as creature-conscious under the standard definition, but which do indeed seem conscious by our intuition. Imagine a machine that has no sensory contact with the outside world (and therefore cannot be said to be aware of its surroundings), and that is always in an unchanging physical state (hence cannot be thought of as awake or asleep), but which constantly experiences a single, static quale (whatever that quale might be). Even though we cannot know that this quale is present—for like all qualia, it would be private and ineffable—I believe most people would consider the presence of such a quale to grant the machine creature consciousness despite its other properties.

The fifth advantage of my definition is that it avoids the obvious charge of circularity that can be leveled at the standard definition as a result of its use of the terms "awake," "aware," and "alert." It is hard to define these terms without use of the notion of consciousness. What is it to be aware of x if not to be conscious of x? We can define the property of being awake through physiological and neural activity, but we do this only

because those aspects of a person are empirically accessible, unlike their qualia. Otherwise, we would probably define awakeness in terms of qualia. The property of alertness is trickier. One might at first be tempted to define it by referring to cognition rather than qualia: I am in a state of alertness if objects and events in my immediate environment affect my thoughts or behavior. But many things affect behavior in instances where alertness is not involved. Differences in air density affect the behavior of an airplane as it flies, but we wouldn't want to claim that the airplane is alert in any significant sense. And unless one is prepared to dismiss completely the idea of the subconscious, it seems clear that the thoughts of a completely non-alert person—say, someone in a deep reverie or on the verge of sleep—can be influenced by other thoughts or perhaps even by external objects or events of which they are unaware. So I think alertness must instead be defined by reference to consciousness: a system is alert if and only if it is creature-conscious. And this raises the specter of circularity if we intend to use alertness in a definition of creature consciousness.

This leaves us with my definition of creature consciousness as the property of having some mental state with qualitative properties—i.e., for a subject to be creature-conscious, it is necessary and sufficient that it have some qualia. On this definition, a person is not creature-conscious if he is deeply anesthetized, comatose, or dead. Dreamless sleep also precludes the experiencing of qualia, and so inhibits creature consciousness. But I see no reason to distinguish the qualia that occur during dreams from the qualia of waking life. It seems arbitrary to discriminate against dream-qualia simply because they are not causally connected to the external world in the same way that waking-qualia are, so I consider dreams to be sufficient for creature consciousness. It is often said that consciousness

exists in degrees, since exhaustion, drugs, injury, emotional shock, or any number of other phenomena can cause a person to be less alert than normal while still being conscious to some degree. For example, Searle likens consciousness to a rheostat: it has distinct on and off positions but an indefinite range of values within the on position.[14] I agree that one can experience degrees of alertness or awareness, but I'm not sure creature consciousness ought to be described in this way. According to my definition of creature consciousness, at any given time you've either got it or you don't, and no meaningful distinctions can be made between different degrees of having it. A single quale, no matter how faint, grants full creature consciousness to a system, and that system is in no way less creature-conscious than a system that enjoys a rich set of varied qualia. I can assert with complete authority the presence only of my own creature consciousness, but it seems overwhelmingly likely that virtually all people are capable of creature consciousness (notable exceptions include those with extraordinarily severe brain injury or misdevelopment, and the permanently comatose). Animals may or may not be capable of creature consciousness, but even those quite low on the phylogenetic scale clearly act as if they are. Physical objects outside of the animal kingdom, such as rose bushes, clouds, and planets almost certainly lack the capacity for creature consciousness. Whether computers can be creature-conscious is a question that consistently provokes strong words and great passion; I'll return to the issue in chapters four and five.

[14] Searle [95, p. 83]

Transitive vs. intransitive creature consciousness

There are two aspects to creature consciousness: transitive and intransitive. Intransitive creature consciousness corresponds what is probably the most common informal use of the term "consciousness." When we describe someone as conscious *simpliciter* (i.e., not conscious *of* something, or being in some conscious *state*), we generally mean that he is awake, alert, and aware of his surroundings. A more precise way to capture what we mean, and a way that avoids the circularity mentioned above, is to say that someone who is intransitively creature conscious is simply experiencing qualia of some sort. Transitive creature consciousness, on the other hand, is the property of being aware of some particular object, event, or perhaps fact. Put another way, transitive creature consciousness is just intransitive creature consciousness with an intentional element added, where by "intentional" I mean the philosophical sense of being directed at something and not the everyday English sense of intending to do something (though the latter is an obvious case of the former). This technical definition of intentionality, and the claim that intentional states do exist, are relatively uncontroversial. What is perhaps less readily agreed upon is how to analyze the core notion of intentionality further. I have found two analyses in particular to be especially helpful, thorough, and persuasive: those presented by Searle [94, ch. 1] and Siewert [99, ch. 6]. These two approaches to intentionality are largely compatible, as each relies heavily on the idea of "fit" between an intentional state and its target. Rather than muddy the waters by presenting yet another interpretation of the concept, I will have these two analyses in mind when I speak of intentionality throughout this dissertation.

Here is an example of transitive vs. intransitive creature consciousness: when I

emerge from a coma or wake after dreamless sleep, I regain intransitive creature consciousness simply by virtue of waking up. At roughly the same time, I become transitively creature-conscious when I become aware of particular things such as the feel of my bed sheets, the tree outside my bedroom window, or the hunger in my gut. Again, a person is *intransitively* creature-conscious if and only if he experiences qualia of any sort. This is the sort of creature consciousness I have been discussing so far. A person is *transitively* creature-conscious if and only if he experiences a quale that is unambiguously associated with a particular object, event, or fact and he possesses an intentional state that is directed at that object, event, or fact. When I look at my red backpack and experience a red quale while having some mental state that points to the backpack (i.e., the mental state has the backpack as its intentional content), that experience qualifies me as being transitively creature-conscious.[15] Notice, then, that there is a cognitive component to transitive creature consciousness apart from the obvious phenomenal component. Not only do I experience the quale (the phenomenal component), but I must have an appropriate intentional state (the cognitive component). In particular, I must be thinking of the object, fact, or event that is responsible for producing the quale in me. Why? Because without this intentional state I am simply experiencing unassociated, free-floating qualia, which is how I have defined intransitive creature consciousness. This cognitive element is all that separates transitive and intransitive aspects of creature consciousness. So to summarize my claim so far: while qualia are an essential part of both transitive and intransitive creature

[15]I will not take a stand on the issue of whether I perceive the backpack directly or instead construct a representation of it from sense data that I immediately perceive. Because both of these interpretations allow for qualia, this agnosticism does not threaten my view of creature consciousness. Nevertheless, this is an important issue in philosophy and especially in psychology, with heavyweights such as William James in the first camp and structuralists in the second.

consciousness, the transitive aspect also requires that the creature-conscious system have a particular intentional state.

I want to emphasize that it is not crucial that the subject *know* that the intentional state and the quale are linked. This would be too strong a condition on transitive creature consciousness, for it would likely exclude any non-human animals, and perhaps other systems as well. It is conceivable that cats, say, are transitively creature conscious (certainly my cat seems conscious of the fact that the doorbell is ringing or that it smells like there is food in his bowl), but it seems unlikely that cats are philosophically sophisticated enough to understand the notion of an intentional state or a quale, much less to know that a particular intentional state is connected in some way to a particular quale. At this stage of analysis we certainly want to leave open the possibility that such animals are transitively creature conscious, so we ought not to impose any restrictions that would exclude such animals.

2.2.2 State consciousness

Let us leave transitive and intransitive creature consciousness and move on to state consciousness. Whereas creature consciousness of either kind is a property of a whole creature, state consciousness is a property of individual mental states. I will stipulate that a mental state is state-conscious if and only if it has at least one qualitative property (i.e., at least one quale is a property of that mental state). As mentioned earlier, I take a mental state to be—expressed informally—one or more element of the contents of a mind at a particular time. I have already emphasized that the difficulties we might have in individuating mental states are, for the purposes of this discussion, unimportant. What is important is that any states that involve qualia are state-conscious. Those that do not, are

not. Note that state consciousness is always intransitive. It is a property of mental states just as yellowness is a property of daisies: neither state consciousness nor yellowness is *of* anything else.

I believe that unconscious mental states exist, though with the possible caveat, as discussed below, that (*pace* Freud) they must be capable of becoming conscious. This means that the mental states that a person has at a given time need not all be state-conscious at that time. In fact, it is extremely unlikely that all would be (or perhaps even could be) simultaneously state-conscious. This is probably for the best, for imagine the mental chaos that would ensue if such a scenario were possible! Instead, it seems that mental states gain and lose state consciousness over time. For example, my mental state of believing that Jakarta is the capital of Indonesia is almost always state-unconscious, though when I am prompted in the right way it can quickly become state-conscious. This is not to say that we never have deliberate control over these transitions; we regularly choose to shift mental states in and out of consciousness (i.e., we either endow them with state consciousness or strip them of state consciousness). I should also note that there are plenty of problem cases for my view: there are states that we just don't know how to describe. For example, sometimes I am aware of the sound of my computer's fan, while other times I am not. One could claim that the fact that I am startled when the fan noise stops suggests that there is some mental state in me all along that tracks the sound coming from the fan, even though that state is not always state-conscious. Alternatively, one could describe this situation by saying that my fan noise-derived mental state was state-conscious all along, but had been shunted to the periphery of my consciousness except at those times when I chose to focus

on it. I don't think anything of real importance rides on how we choose to describe these odd cases, so long as we realize that there are at least some mental states that have no state consciousness at all at a given time.

My position on the possibility of state-unconscious mental states has some support in the literature. Tyler Burge agrees that mental states need not always be conscious, though he goes on to make the extraordinarily bold (and according to my definition of consciousness, blatantly false) claim that even *phenomenal* mental states (i.e., those that essentially involve qualia) need not be conscious:

> [I]t may be that there is no feeling of a pain (a pain that is nevertheless present) for the hypnotic. There may be no way that it is actually and occurrently like for the epileptic to drive the car. Perhaps what enables the epileptic to operate is the phenomenally registered information (which has phenomenal what-it-is-like qualities that are simply not actually felt by the individual), operating through limited unconscious mental and sensory-motor procedures. But the phenomenal states and the information might be in no sense conscious. [13, p. 433]

This extremely strong claim is debatable. Equally debatable is the question of whether a mental state that never reaches state consciousness can be considered a true mental state at all. Searle's connection principle suggests that such the possibility of state consciousness is a prerequisite for something being a mental state at all,[16] while others have argued that some mental states can remain unconscious throughout our lives.[17] I'm inclined to think of permanently qualia-less, hence unconscious, states as being physical properties of the brain than true psychological entities, but I will officially remain neutral on this subject. This should not affect any of my claims.

[16]See Searle [95, ch. 7] for a discussion of his connection principle.

[17]See Freud [35, chs. 1–3]. Lashley [54] gives a more ambiguous, if no less vigorous, argument for what seems to be the same position.

An example to illustrate the relation between qualia and state consciousness is in order. Imagine a woman eating lunch while having an animated discussion with another person. The woman may be ravenous; great hunger may be one of her mental states at the time. Imagine further that she is not aware of this mental state until her dining companion points out that she is eating faster than usual. This realization is not enough to make her hunger become state-conscious, for until she turns her attention from her conversation and actually feels the hunger (i.e., experiences the qualia associated with it), she will merely be aware of the fact that she is hungry rather than conscious of the hunger itself. This is not sufficient for state consciousness. She must feel the hunger—she must experience that gnawing, unpleasant sensation in order for her hunger to become state-conscious.

One might be tempted to describe this situation by saying that the woman's hunger started at the *periphery* of her consciousness and then moved to the *center* of her consciousness as she became aware of it. This is an appropriate way to understand the shifting degree of hunger that she feels as her attention moves toward or away from that hunger, but I think transition from periphery to center can only occur after the state of hunger has moved from being state-unconscious to being state-conscious. When the hunger is fully state-unconscious, the woman feels no qualia associated with it at all. It may then become state-conscious by entering the periphery of her consciousness, at which point it is a weak hunger that she is only vaguely aware of. Finally, as she turns her attention toward the hunger and feels it more acutely, it enters the center of her consciousness. So while the center vs. periphery distinction is a valid and helpful distinction to make, it should not be confused with the state-conscious vs. state-unconscious distinction. Only state-conscious

states can be in the center or periphery of consciousness; it makes no sense to speak of state-unconscious states being in either the center or the periphery of consciousness.

2.2.3 Relations between the three aspects of consciousness

The three aspects of consciousness I have discussed so far—intransitive creature consciousness, transitive creature consciousness, and state consciousness—are very closely linked. Although nothing of any great importance hinges on the nature of the relations between them, I think investigating these relations will help us better understand the concepts involved. But first, a quick review of definitions.

- A person is *intransitively creature-conscious* at time t iff at time t he has at least one mental state with a qualitative feature or property.
- A person is *transitively creature-conscious* at time t iff at time t he has at least one mental state that has both qualitative properties and intentional content.
- A mental state is *state-conscious* at time t if and only if it has at least one qualitative property at time t.

Now on to the relations among these concepts. Let's start with intransitive creature consciousness. One might be tempted to think that a person cannot be intransitively creature-conscious unless he is simultaneously transitively creature-conscious. In order to be conscious, mustn't one be conscious *of* something? I claim that this is not the case. Imagine that Smith is intransitively creature-conscious in virtue of experiencing the quale of, say, anxiety. He need not be anxious about anything in particular. Indeed, panic attacks are well-known phenomena that are sometimes described as moments of anxiety that

aren't attached to any object, fact, or event. This suggests that Smith can be intransitively creature-conscious without being transitively creature-conscious of anything. But mustn't he at least be transitively creature-conscious of the fact that he is feeling anxious? In other words, wouldn't he have to have a state-conscious mental state representing the fact that he is anxious? Or to pose the question more casually: wouldn't an anxious person have to know that he is anxious? I don't think so. It seems possible to imagine a person who feels anxious, but who is so distracted by some task or by some stronger feeling that he does not become aware of the fact that he is anxious. He may remain this way until the anxiety passes, or until someone else points out that he is acting strangely and suggests that he may be anxious about something. Or imagine someone who is extraordinarily inattentive to his own mental states. Perhaps he is so intently focused on other people that he never gets around to introspecting, and so remains ignorant of his own anxiety. This is conceivable even in cases where the qualia associated with the anxiety are quite powerful. Or for a more *outré* example of intransitive creature consciousness without transitive creature consciousness, consider the reputedly cognitively empty state one is said to arrive at during sustained Buddhist meditation. I don't want to rely too heavily on this example since it might be argued that people in deep meditative trances lose intransitive creature consciousness as well, and because I think the anxiety example is sufficient to establish my point. However, I will concede that intransitive creature consciousness is almost always accompanied by transitive creature consciousness. States that have qualitative properties frequently have intentional content that point to the objects, facts, or events that are linked to those qualia, and so make the bearer of those states transitively creature-conscious as well.

I realize my position will strike many as counterintuitive—how can one be conscious while not being conscious *of* anything?—so I will provide one more example to strengthen my case. Say I gaze absent-mindedly at the cloudless sky and see an expanse of blue. I am intransitively creature conscious in virtue of experiencing this blue quale. I also have a state-conscious mental state that represents this blue field. But if I am thinking intently of other things at the time, I might pay no attention to the fact that I am looking at the sky, and I might not consider the fact that I am seeing a pure blue field. The mental state caused by my sense perception would have only phenomenal, and not intentional, content. As such, it would incapable of being right or wrong, accurate or inaccurate, satisfied or unsatisfied. It would consist of pure, undirected qualia that are entirely free of intentionality. In this case, I would not be transitively creature-conscious of the sky or of the fact that I am looking at the sky. If someone were then to interrupt my reverie by asking what I am looking at, I would likely focus my attention on my visual experience, realize that I am looking at the sky, and thereby become transitively creature-conscious both of the sky and of the fact that I am looking at it.[18]

One might object that this example is able to eliminate transitive consciousness of the sky only by replacing it with transitive consciousness of something else—of whatever it is that distracts me sufficiently to keep me from thinking about the sky. But I think we can imagine cases where I lack transitive consciousness not just of the sky, but of any sort

[18]Note that this is not quite the same as saying that my visual experience of the sky would move from the periphery to the center of my conscious field. I take it that having something at the periphery of one's consciousness is essentially a weaker form of having that thing at the center of one's consciousness: you're still aware of the thing to a certain (lesser) degree, but you're not thinking much about it or paying much attention to it. But the sky-gazing example I am describing is not just a case of *weak* awareness that one is looking at the sky, but is instead a case of a complete *lack* of awareness that one is looking at the sky. *Any* degree of awareness about what exactly he is looking at would make our sky-gazer transitively creature-conscious.

at all. Cases where I am on the verge of sleep may be like this, as are cases where I'm thinking of nothing in particular. Most of us have experienced times when our minds are simply idle, where sensory perceptions do impinge on our consciousness somewhat, but we don't think about or do anything with those perceptions. I maintain that these are cases in which we are intransitively conscious without being transitively conscious of anything.

Dretske [29] makes a similar claim about intransitive creature consciousness not requiring transitive creature consciousness when he notes that in the case of vivid hallucination it seems appropriate to describe a person as intransitively creature-conscious but not transitively creature-conscious of anything that actually exists, hence not transitively creature-conscious at all. This suggests that intransitive creature consciousness can exist without transitive creature consciousness. One might be tempted to argue with him by saying that we routinely use language to refer to non-existent things and events (e.g., Santa Claus, the 2.2-child family, the burning of Tara), and since we can speak of or have other intentional states directed at these things, then we ought to be able to be transitively creature conscious of non-existent things as well. But this move is unwise. Any intentional state I have about Sherlock Holmes is false in an important sense. To borrow more terminology from Searle [94, ch. 1], the conditions of satisfaction of such an intentional state are not satisfied, and could never be, as the object of that intentional state does not exist. An intentional state directed at a real object (say, my belief that my grandfather's Hampden pocket watch is sitting on my desk) can be satisfied if the real world corresponds to its conditions of satisfaction (in this case, if that very watch is actually sitting on my desk). But an intentional state directed at an imaginary object can only be satisfied in an imaginary sense.

We can only *pretend* that it is satisfied, for the real world is missing the objects required to *actually* satisfy its conditions of satisfaction. We are only play-acting when we refer to Sherlock Holmes, and so we can only play-act at being conscious of Sherlock Holmes.[19] Note, however, that this restriction holds only for *transitive* creature consciousness. The qualia-based definition of creature consciousness that I have provided suggests that what is essential to *intransitive* creature consciousness is not that the qualia that produce it hook onto or reflect the real world in any way, but simply that they register on the mind of a person. So a hallucination would suffice to sustain intransitive creature consciousness, if not transitive creature consciousness.

I have just claimed that intransitive creature consciousness does not require concomitant transitive creature consciousness. But I think intransitive creature consciousness does always require at least one state-conscious mental state. In the case of Smith's aforementioned free-floating anxiety, he is not only intransitively creature-conscious in virtue of experiencing the qualia that regrettably accompany anxiety, but he also has the state-conscious mental state of anxiety. In addition, he will sometimes have the state-conscious mental state of knowing that he is anxious, although the arrival of this state will also signal the arrival of transitive creature consciousness: he becomes transitively conscious of the fact that he is anxious. Speaking more generally, the experiencing of quale x not only grants him intransitive creature consciousness, but must also produce in him the state-conscious state of experiencing quale x and will sometimes also produce in him the state-conscious state of knowing that he is experiencing quale x. So each instance of intransitive creature

[19]The ontology of fictional entities is a fascinating branch of metaphysics. By my lights, some of the best work in this area has come from Kendall Walton [112].

consciousness must be accompanied by at least one state-conscious mental state.

Now let's look at transitive creature consciousness. It seems to require both of the other aspects of consciousness. Say that Jones is transitively creature-conscious of the fact that he is wearing shoes that are too small. This means that he experiences qualia produced by this fact—perhaps pain, pressure, and stiffness in his feet and toes—and he has an intentional state that is directed at that fact. First, it seems clear that this would require that he be intransitively creature-conscious (i.e., he must be experiencing *some* quale). While this claim that intransitive creature consciousness is a necessary condition for transitive creature consciousness has garnered a fair amount of support in the literature,[20] it is not completely without its critics. For example, Dretske [29, footnote 12] suggests dreaming as a possible counterexample to this claim: a dreaming person would seem to be transitively creature-conscious but perhaps not intransitively creature-conscious.[21] This is an important case, but I think it reveals a different truth than Dretske intends. While he uses it to argue that transitive creature consciousness can occur without intransitive creature consciousness, I think it makes a better case for the idea that being awake is not a necessary condition of intransitive creature consciousness. The experiencing of qualia can occur when a subject is either awake or asleep. Intransitive creature consciousness does not require that qualia accurately represent the world, so the fact that dream-qualia often provide a false picture of the world does not prevent their presence from granting intransitive creature consciousness to their subject.

[20]For example, White [116, p. 59]: "If there is anything of which a man is conscious, it follows that he is conscious; to lose consciousness is to cease to be conscious of anything."

[21]Locke [58, p. 336] appears to make an even stronger claim about sleep: "in sound sleep having no thoughts at all, or at least none with that consciousness which remarks our waking thoughts...." If Locke is referring to all sleep and not just dreamless sleep, then this would seem to preclude even transitive creature consciousness during sleep of any kind. This position is surely wrong.

My claim is supported by evidence from a lucid dreaming experiment run by Stanford psychologist Stephen LaBerge [53]. Lucid dreaming occurs when a subject knows that he is asleep and dreaming, and in his dreaming state is able to follow instructions provided when he was awake. The details of the experiment are as follows. LaBerge first had the fully awake subject visually track the tip of his right index finger as he used this finger to trace a circle in the air in front of his body. This perceptual activity produced slow, smooth eye movements in all subjects. Then LaBerge had the subject close his eyes and imagine tracing the same circle while actually tracking the imaginary movement of his finger with his closed eyes. This exercise in imagination produced rapid, jerky, saccadic eye movements in all subjects. When these subjects later traced circles and tracked finger movements in lucid dreams, the resulting physical eye movement was slow and smooth—virtually identical to the movement produced during the perceptual experiment. LaBerge takes this to indicate that the neurological processes involved in dreaming are much more similar to those invoked during perception than to those involved in cases of imagination. How does this support my argument against Dretske? I have already stated that qualia are an essential part of any perception, and if I have done a thorough job of explaining what it means to experience a quale, then it should be just as clear that qualia unambiguously attend vivid (and even not-so-vivid) episodes of imagination. This means that imagination cannot occur without intransitive creature consciousness. Even without the results of LaBerge's experiment, this might give us reason to think that dreaming perception likewise requires intransitive creature consciousness, for at least in my own experience dreams are frequently even more vivid—more qualia-laden—than episodes of waking imagination. But if we are

given evidence that the perception involved in dreams is neurologically closer to waking perception than to waking imagination, then we have an even stronger reason to think qualia are experienced during dreams, which means that dreams quite clearly require intransitive creature consciousness. So Dretske is wrong to think that dreams are a case of transitive creature consciousness without concurrent intransitive creature consciousness.

David Armstrong [2] presents a famous scenario which could be taken as a counterexample to the claim I have just made. Adverting to an experience many have had, he describes a tired driver who mentally fades out while on the road, only to come to several minutes and several miles later. Clearly the driver was conscious in some sense during that span, since a fully unconscious person would have had no chance of keeping the car on the road. But he was just as clearly lacking some form of consciousness that a normal, alert driver has. This example could be thought of as a case in which a person is transitively creature-conscious without being intransitively creature-conscious: the driver in some sense "saw" the road, the stop signs, and the other traffic, while he himself was not conscious. But I think this is not the only way we can describe the situation. It makes just as much sense to say that the driver was fully creature-conscious in both a transitive and an intransitive sense, but that he simply did not form any long-term memories of the state-conscious states that represented the road, stop signs, and other traffic. On this reading, the example does not threaten my claim that transitive creature consciousness requires intransitive creature consciousness.

Let us return to the example of the man who is transitively creature-conscious of the fact that his shoes are too small. I claim that he must not only be intransitively

creature-conscious (as just discussed), but that he must also have the state-conscious mental state of feeling pain, pressure, and stiffness in his feet and toes. Furthermore, he will have the state-conscious mental state of knowing that he is wearing shoes that are too small.[22] Thus transitive creature consciousness will always have as necessary conditions both intransitive creature consciousness and state consciousness. To experience a quale produced by a particular thing and to have an intentional state representing the cause of the quale (i.e., to be transitively creature-conscious of that thing), one must both be capable of experiencing qualia in general (i.e., be intransitively creature-conscious) and actually have a mental state that has that quale as a property (i.e., have an appropriate state-conscious mental state).

Finally, let us turn to state consciousness. By my definition of state consciousness, for Smith to have a state-conscious mental state representing an apple in front of him he must both have a mental state that represents the apple, and experience qualia that are properties of that state. Regardless of whether the content of the mental state is phenomenal in character (he sees the apple, and sees that it is in front of him) or propositional (with his eyes closed, he is told that there is an apple in front of him and he forms the thought that this is so), there will be qualia associated with the mental state. In the phenomenal case, he experiences qualia of redness, roundness, and whatever else is visually distinctive about an apple. In the propositional case, he experiences qualia that are associated with the thought that there is an apple in front of him. Smith must be intransitively creature-conscious in either case, for he must be capable of experiencing qualia in general. He also must be

[22] I have already issued a promissory note regarding my argument that at least some propositional knowledge of this sort involves qualia; I will make good on the note in chapter three.

transitively creature-conscious. This is because in order for his state-conscious mental state to represent the apple, it must be intentional: it must point to the apple and be about the apple in some sense. When this intentional state is combined with the qualia produced by the apple, Smith becomes transitively creature-conscious of the apple. Furthermore, he most likely will also become transitively creature-conscious of the fact that there is an apple in front of him, although this is not required for state consciousness. Even if he fails to see the apple *as an apple*—say he perceives a round, shiny, red blob but for some reason does not realize that it is an apple—he still must be transitively creature-conscious of whatever it is he does see. In this case, he would be transitively creature-conscious of the round, shiny, red blob and of the fact that it is in front of him. So state-conscious mental states are always accompanied by transitive creature consciousness.

So far I have only discussed state consciousness vis-à-vis representational states such as seeing an apple and believing that it is in front of you. Can non-representational states also be state-conscious? I believe so. To return to two earlier examples, imagine either feeling free-floating anxiety or else having your visual field taken up entirely by a single shade of blue, such as occurs when you stare at the cloudless sky. I maintain that the mental states that make up either being anxious or seeing a vast expanse of monochromatic blue are non-representational—what could they represent?—but are nevertheless state conscious in virtue of having qualitative properties. Such cases are rare, and certainly the vast majority of state conscious states are representational, hence are accompanied by transitive creature consciousness. But transitive creature consciousness is not a necessary condition on state consciousness.

We have covered a lot of ground, and a quick summary of the interrelations of the three aspects of consciousness is in order. My claims are these:

- *Intransitive creature consciousness* has one necessary and sufficient condition: the subject must have at least one state-conscious mental state. That is, he must have at least one mental state that has at least one qualitative property. He need not be transitively creature-conscious of anything.

- *Transitive creature consciousness* has as necessary and jointly sufficient conditions both intransitive creature consciousness and the possession of at least one state-conscious mental state.

- Possession of a *state-conscious* mental state has intransitive creature consciousness as a necessary and sufficient condition, though it very rarely occurs without transitive creature consciousness as well.

Although one of my goals in this chapter has been to make clear the differences between these three aspects of the single phenomenon that is consciousness, I freely admit that the conceptual boundaries are blurry enough to sometimes cause confusion. This confusion is easy to bring about: introspect for a moment, and "look" carefully at the state of your own mind and try to figure out what sorts of consciousness are occurring. If your experience is anything like mine, you will find it extraordinarily difficult to know whether you are "looking" at the mental states of which you are currently aware (i.e., you are observing your state-conscious mental states), at your awareness of mental states (i.e., you are observing your own transitive creature consciousness), or at the general, non-target-specific

property of awareness that you have (i.e., you are observing your own intransitive creature consciousness). To confuse matters further, it is quite possible that your introspection has looked right through these aspects of consciousness and has settled instead on the objects or events in the world which you currently perceive. My point is not that "state consciousness," "transitive creature consciousness," and "intransitive creature consciousness" are just different terms for the same thing, but rather that the aspects of consciousness that they point to are closely related—so closely, in fact, that it sometimes can be very difficult to pull them apart or to know which facet of consciousness a particular comment applies to. Some philosophers who have recognized the differences between these different aspects of consciousness have found it more helpful to conflate them rather than to continue to treat them as distinct.[23] But while the tightness of these interrelations can be confusing, they need not trouble us. It may well be that many comments on consciousness apply not to any particular feature but rather to the general phenomenon. Even when these three aspects can be separated, discussion of any one will inevitably involve discussion of the others as well.[24]

Nevertheless, the rest of this dissertation will concentrate—to the extent that it is possible to focus on one aspect to the exclusion of the other two—on intransitive creature consciousness, and use of the term "consciousness" simpliciter should be understood as referring to that facet of the overall phenomenon. Whenever I do need to make comments about one of the other two aspects, I will try to make clear exactly which aspect I am

[23]Stubenberg [104] takes this approach, despite being extraordinarily sensitive to the differences between these facets of consciousness.

[24]For a very different take on the interrelations between these aspects see Rosenthal [86]. He agrees for the most part with the tripartite distinction I have presented, but claims that these are really names for three different *forms* of consciousness, each of which is largely independent of the other two—so independent, in fact, that unraveling the mysteries of one will shed little or no light on the others.

speaking of. Some philosophers have argued that state consciousness is the core notion of consciousness, and that any other aspects derive from this.[25] Those who subscribe to that view may be disappointed at my focus on intransitive creature consciousness. To them I can only respond that although the issue of state consciousness is a fascinating one, in order to treat the topic properly one must first provide an extensive analysis of the ontological status and structure of mental states and of the relations between them. That task, while certainly worthwhile, lies well beyond the scope of this work.

[25]For example, Goldman [37].

Chapter 3

Filling in some gaps

I have so far left some conspicuous lacunae in my characterization of the nature of consciousness. My aim in this chapter is to fill these to some degree. First, I will argue that many instances of propositional thought involve qualia just as real and just as vivid as the qualia that are associated with sense perception. This is important for me to establish, for it lends crucial support to my claim that consciousness is a pervasive aspect of mental life. Second, I will discuss alternative notions of consciousness that have been proposed by other philosophers. I will argue that none of these definitions threatens the legitimacy of my own, for each either reduces to a version of my definition or else presents significant difficulties not faced by my concept of consciousness.

3.1 Qualia and propositional thought

> *When I think of a lion ... there seems to be a whiff of leonine quality to my phenomenology: what it is like to think of a lion is subtly different from what it is like to think of the Eiffel tower. More obviously, cognitive attitudes such as desire often have a strong phenomenal flavor. Desire seems to exert a phenomenological "tug," and memory often has*

a qualitative component, as with the experience of nostalgia or regret.

David Chalmers [15, p. 10]

I promised earlier to show that at least some of our propositional thoughts involve qualia, and I will now make good on that promise. The core of my argument for this view is almost embarrassingly simple: as Chalmers suggests in the quotation above, all that we need to do to see that propositional thoughts have qualia is to call up two different propositional thoughts and pay attention to the fact that it just plain *feels different* to think the first of these thoughts than it does to think the second. Try this even with two closely related thoughts, such as the thought "Portland is north of Boston" and the thought "Boston is north of New York." Or better yet, take the same concept expressed in two different ways: "Harvard is older than Yale" vs. "Yale is younger than Harvard." Unless your experience is radically different from mine, you will immediately notice that these thoughts all involve quite different qualia.

Here is another argument for the same claim. We often have strong occurrent beliefs and desires that are not actually put into words. Yesterday around noon I felt hungry, I believed that I was hungry, and I wanted to act in hunger-relieving ways. But at no time yesterday did I actually string together the words "I am hungry"—I neither spoke nor explicitly thought these words. Yet it seems clear (to me, at least, from my first-person vantage point) that I did have some sort of hunger-related propositional thought at that time. Now what could this thought have consisted in if not words in English or mentalese? Qualia are the only candidate components left, so I submit that the thought must have

been purely qualitative.[1]

But why have I hedged by saying that only *some* of our propositional thoughts involve qualia? Well, it is often said that we have unconscious as well as conscious propositional thoughts, and as we well know by now, my definition of consciousness dictates that unconscious mental states have no qualia associated with them. For example, it might be proper to say that I believe that the Dalmatian coast borders the Adriatic Sea even when I am not thinking about Croatian geography. This belief, which would presumably qualify as an unconscious belief, would have propositional content but no qualia. To take a more extreme case, it would be very strange to say (though I'm sure there are philosophers who would say exactly this) that I have no beliefs whatsoever when I am in dreamless sleep or am comatose. Since beliefs are a type of propositional thought, it seems appropriate to allow for the possibility of qualia-less propositional thoughts, with the caveat that any such thoughts would by definition have to be unconscious (hence not terribly interesting). But our standard, run-of-the-mill thoughts, beliefs, desires, and other propositional mental states are both state-conscious and (as follows from my definition of state-consciousness) qualia-laden.

I realize these arguments are far from airtight, as they rely heavily on analysis of my own experience with my own mental states. But I have no reason to think that other people experience their states in relevantly different ways than I do, so I feel justified

[1]I should at least mention one further argument for this point. It is sometimes suggested that an easy way to see that propositional thought has qualitative features is to consider the same propositional content expressed in different languages. Apparently it feels very different to think "I am hungry" than it does to think "J'ai faim," "Yo tengo hambre," or (if you speak Gujarati) "Mane bhuk lage che." This argument sounds entirely plausible, but I am not sufficiently proficient in any language other than English to test it myself; the difference in comfort and naturalness between any English phrase and the equivalent phrase in another language swamps any qualitative differences that I might feel.

in assuming that their propositional thoughts are just as qualia-rich as mine are. In the absence of any further arguments toward this point, let us move on to the next gap in my account of consciousness that needs to be filled: I need to explain why many standard accounts of consciousness are wrong.

3.2 What consciousness isn't

> *I believe the notion of phenomenal consciousness is the core notion of consciousness. Any being that is not phenomenally conscious is not conscious in any sense. There are no zombies that lack phenomenal consciousness but are conscious in some further way.... I do not know how to defend this view. I do not know why it is true. But despite a literature replete with assumptions to the contrary, I find it compelling.*
>
> Tyler Burge [13, pp. 428–9]

Like Burge, I have suggested that one is conscious if and only if one is experiencing qualia. This focus is not unique in the literature; a number of philosophers have zeroed in on qualia as the core notion of consciousness. Chalmers, for example, agrees that getting a full understanding of qualia is the central task in unraveling the mysteries of consciousness. He emphasizes that even after we have solved some of the "easy" psychological or behavioral problems mentioned in chapter one, we are still left with the more mysterious issue of why qualia occur at all, why they are only produced (as far as we know) by certain brain processes, and why *those* qualia and not others accompany certain brain processes. This gap is particularly evident in the case of phenomenal judgments, such as when a person who is looking at a tomato is asked to introspect in such a way as to reveal the phenomenal qualities of his occurrent mental states, and he responds by saying or thinking "I see a round, red splotch in front of me." Even after we have explained why he made that judgment about

that experiences or mental state (i.e., once our psychological theories can account for his behavior), we still need to explain why seeing the tomato *seemed* a certain way to him. This is ultimately the most interesting question we must address when dealing with consciousness, and is probably the last consciousness-related problem we will be able to crack. Chalmers, for one, agrees:

> [W]hen one starts from phenomenal judgments as the explananda of one's theory of consciousness, one will inevitably be led to a reductive view. But the ultimate explananda are not the judgments but experiences themselves. No mere explanation of dispositions to behave will explain why there is something it is like to be a conscious agent. [15, p. 191]

Flanagan also points to qualia as the central concept within the constellation of phenomena and activities that is sometimes thought to constitute consciousness. He takes pains to emphasize that a wide variety of mental states and processes can be accompanied by qualia: "[C]onsciousness [is] a name for a heterogeneous set of events and processes that share *the property of being experienced*. Consciousness is taken to name a set of processes, not a thing or a mental faculty." [emphasis mine] [33, p. 220] Lycan locates the heart of consciousness in exactly the same place: "What really concern me are the qualitative features or phenomenal characters of mental items." [60, p. xii] And Searle gives perhaps the most thorough statement of the same position when he suggests that qualia are ubiquitous—they are features of any conscious mental state:

> Conscious states are qualitative in the sense that for any conscious state, such as feeling a pain or worrying about the economic situation, there is something that it qualitatively feels like to be in that state.... [A]ll conscious phenomena are qualitative, subjective experiences, and hence are qualia. There are not two types of phenomena, consciousness and qualia. There is just consciousness, which is a series of qualitative states. [97, pp. xiv, 9]

Other philosophers introduce more ambiguous definitions of consciousness that are not, strictly speaking, contrary to my definition. Armstrong [2] presents an excellent example: he asserts that a mind is minimally creature-conscious at time t if and only if there is any mental activity occurring in that mind at time t. Certainly the experiencing of qualia, whether from immediate perception or recalled memories, would count as mental activity. Presumably propositional thoughts would qualify, as would the experiencing of emotions or moods. Indeed, any sort of human cognition I can think of all seem to be accompanied by qualia. If this is true, then Armstrong's definition of consciousness is consonant with mine.

It is sometimes the case that definitions of consciousness that prima facie seem utterly contrary to my qualia-based definition can be shown, on closer examination, actually to be compatible with my view. Consider psychologist Bernard Baars, who in much of his recent writing has defined consciousness in terms of *attention* [5]. He contrasts conscious mental phenomena with unconscious but otherwise equivalent phenomena, claiming that the two differ only in that the former are attended to while the latter occur automatically in some sense. Attended streams of stimulation are conscious, and previously unattended events that interrupt the attended stream become conscious. In general, when attention is drawn to a particular stimulus, the subject becomes conscious of that stimulus. Some of the implications that he draws from this view would perhaps not obviously follow from my definition of consciousness (for example, he notes that unattended perceptual channels seem not to produce any learning), but in the main his theory of consciousness meshes nicely with my own, probably due to the fact that states or processes that are automatic or unattended are also qualia-free. In saying that a state that is unattended has no qualitative properties,

and that such a state can only be unconscious, Baars seems to be tacitly agreeing with my claim that qualia are essential properties of any conscious state.

At first blush, Alvin Goldman also appears to give a different definition than I do, but upon further inspection our positions do not seem so very far apart. He starts off on the right foot by making the same distinction between state, intransitive creature, and transitive creature consciousness that I do. But he then proclaims state consciousness to be the core notion, and defines the other two aspects in terms of conscious states. However, our positions are reconcilable once we realize that his definition of a conscious state is built around the notion of qualia: "A partial psychological state is conscious if and only if it involves phenomenal awareness, that is, subjective experience or feeling." [37, p. 364] Furthermore, his account of creature consciousness draws on qualia (via his definition of state consciousness) no less than does my own. We also agree that an intransitively creature-conscious person must be in some conscious mental state or other: "Given the notion of a conscious partial state, we can say that a person's generalized condition at a given moment is conscious [i.e., he is intransitively creature-conscious] iff he possesses at least one conscious partial state at that time." [37, p. 364].

But there are a raft of phenomena or concepts other than qualia that are sometimes associated with consciousness, and in my opinion, wrongly so. I will discuss some of these now, showing how they differ from my view. At the end of this section, I will argue that none of these is a helpful definition, and that the phenomena they are based upon should be excluded from consciousness in its purest sense.

3.2.1 Consciousness is not wakefulness

We will start with an easy one. It is sometimes suggested that consciousness is just another term for *wakefulness*: anyone who is awake is conscious, and one cannot be conscious without being awake.[2] I have already argued explicitly against the second of these claims by saying that qualia-filled dreams produce intransitive creature consciousness just as surely as any waking experiences do. The context within which dream qualia occur can seem strange, to be sure: a dream of a purple elephant speaking to me in Dutch would likely seem surreal even while I was dreaming, but the qualitative components of that dream—the purpleness and smell of the animal and the sound of its speech—would be just as real as if I had experienced them while awake. No one would dispute that I am in some mental state or other while experiencing this dream, and while the qualitative properties of that mental state might be unusual or unexpected, they are no less real than the qualitative properties of any mental state I might be in while awake. So I think it is quite clear that wakefulness is not a necessary condition for experiencing qualia, and hence is not a necessary condition for consciousness.

But the first claim made above is probably true: it *does* seem that anyone who is awake is conscious. This intuition rests on another claim that is, admittedly, potentially defeasible. The second claim is this: it seems likely that anyone who is awake must be in some mental state or other. To see this, think of what it would be like to be awake but not

[2]Searle is sometimes cited as holding this position, due to his definition of consciousness: "When I wake up from a dreamless sleep, I enter a state of consciousness, a state that continues as long as I am awake. When I go to sleep or am put under a general anesthetic or die, my conscious states cease." [95, p. 83] But this is to misunderstand his position, for he also allows for consciousness during *dreaming* sleep: "If during sleep I have dreams, I become conscious, though dream forms of consciousness in general are of a much lower intensity and vividness than ordinary waking consciousness." [95, p. 83]]

be in any mental state. One's mind would have to be completely blank: devoid of thoughts, emotions, images, memories, or perceptual input of any kind. Perhaps others have encountered this sort of mental emptiness—it almost sounds blissful, though the feeling of bliss would immediately introduce qualia into the scenario—but I am fairly certain that I have not. However, I do not want to sound too dogmatic on this point considering that it would be hard to know whether one had spent any time without being in any mental state at all, given that there would be nothing distinctive about that stretch of time that would allow it to be remembered later! It might be suggested that the subject could be told by other observers that he appeared to be awake though mentally vacant, but it is hard to see why such a subject would qualify as being awake. So despite the possibility of counterexamples (from practitioners of meditative religions, say), I feel fairly comfortable saying that awakeness requires that the subject be in some mental state, and that that mental state must have qualitative properties—i.e., being in a single state-unconscious mental state is not enough to qualify someone as being awake, unless the term "awake" is being used in some truly unorthodox sense (and it is hard even to imagine what such a sense might be).

So anyone who is awake is indeed conscious. But this is not to say that awakeness is synonymous with consciousness. It is a sufficient condition on consciousness, but, as the dreaming example demonstrates, hardly a necessary one. So we can dismiss the view that consciousness can simply be defined as wakefulness.

3.2.2 Consciousness is not awareness

When consciousness is defined in terms of some other concept, that concept is most often awareness.[3] But I think it is unwise to equate the two phenomena. If I say that Mary is aware of the cat sleeping on her lap, or that Jim is aware that there is a car speeding toward him, I mean that these people have knowledge of a particular object or fact, respectively. In fact, we can simplify this definition once we see that Mary could just as well be said to be aware of the fact that there is a cat sleeping on her lap. I think we lose nothing by making this translation in all cases of awareness.[4] So every case of awareness can appropriately be described as nothing more than knowledge of a fact. It is very important to emphasize this definition of awareness in order to distinguish it from another usage of the term that is sometimes seen in consciousness-related literature, where "to be aware of x" is used synonymously with "to have a phenomenal experience (i.e., qualia) of or caused by x."[5] This second usage of the term obviously makes the property of awareness identical to the property of consciousness. But the more interesting issue is whether awareness is the same thing as consciousness on my (what I hope to be more common-sensical) definition of awareness. Let us address this issue in two parts. First, is consciousness a necessary condition for awareness? Second, is it sufficient?

[3] For example, see Dretske [28, 29].

[4] Searle makes a similar point when he claims that all perception is really "perception-that," or perception *that something is the case* rather than mere perception *of some object* [94, pp. 40–42].

[5] Alvin Goldman is a good example of someone who uses the term in the way that I am discouraging: "a partial psychological state is conscious if and only if it involves phenomenal awareness, that is, subjective experience or feeling." [37, p. 364] Clearly we agree in our ultimate definitions of consciousness, even if we use different terms to express the concepts involved.

Is consciousness a necessary condition for awareness?

If we consider awareness just to be a form of propositional thought, as I have suggested, then it does seem to require consciousness. Think again of Jim, who sees a car speeding straight toward him and is aware of this fact. Imagine that he writes "there is a car coming toward this piece of paper" on a piece of paper and sets it in front of him. That paper clearly is not aware of the car's path, while Jim just as clearly *is* aware of the car. What drives us to distinguish the paper's lack of awareness from Jim's awareness? I think it must be that the paper just isn't capable of propositional thought—it hasn't the right sort of physical makeup to think of anything at all. I can't prove this, and indeed panpsychists would not be at all happy with a claim of this sort, but I trust the rest of my audience will find this view entirely uncontroversial. Lacking the capacity for this sort of thought, a simple piece of paper simply cannot sustain any awareness. But what about more complex but still non-conscious systems like computers? I will also discuss the potential for computer consciousness in chapter four, but for now let us assume that the computer under discussion is certifiably non-conscious (imagine that we have disconnected its consciousness circuit, if you need help thinking of computers as fully non-conscious). Now, couldn't a nonconscious computer become aware of the car were we to toggle on the "car speeding directly toward me" register in its memory? Of course not, and for the same reason that the piece of paper was not aware. In both cases we have systems that represent a fact, but in both cases this representation is imposed from outside by conscious agents. Neither the words on the paper nor the register in the computer intrinsically represent anything about a speeding car; the squiggles of ink and patterns of voltage mean what we say they do *simply because* we say

that they do so.

To make this point even more bluntly, imagine a flat rock with a smooth side and a rough side. Jim declares that when the rough side faces up, it means there is a car speeding toward the rock. Having the smooth side face up means that that there is no such car. We clearly don't want to say that the rock is aware of the presence of any nearby cars, even when its rough side is up. This is because it is merely being used to *represent* an awareness that occurs in an entirely different system (i.e., Jim). A computer, though enormously more complicated than this rock, can have at most the same pseudo-awareness as the rock. In neither the rock nor the computer is there actual propositional thought, and so in neither case is there qualia, and so in neither case is there awareness. This argument of course relies on the claim made earlier in this chapter that conscious propositional thought must be accompanied by qualia. Once we accept that claim, we see that external awareness must also be accompanied by qualia. Since the experiencing of qualia is therefore a necessary condition for awareness, and consciousness just is the experiencing of qualia, we see that consciousness is a necessary condition for awareness. Without consciousness, what appears to be awareness is merely a proxy for the awareness intrinsic to another system.

It may be helpful to restate this point in terms of intentionality rather than propositional thought. What I have claimed so far could just as well be expressed by saying that pieces of paper, rocks, and computers are not capable of having intentional states directed at speeding cars, whereas Jim is fully capable of having such states. Any states in the first three systems that *appear* to be intentional actually involve only *derived* intentionality rather than honest-to-goodness, full-fledged *intrinsic* intentionality.[6] Words written on a

[6]This terminology is not original: see Searle [95, ch. 3].

page, a particular physical orientation of a rock, and data in a computer all can represent things in the world, but they do so only by relying on other intrinsically intentional systems (such as people) from which they derive their intentionality. Without an intrinsic intentional state directed at x, a system cannot properly be said to be aware of x. And because intentional states must involve qualia (as I argued earlier in this chapter), intentional systems must be conscious. So systems incapable of consciousness are incapable of being aware.

However, the psychological literature is replete with cases of a phenomenon known as *blindsight*, which suggest that I am wrong to claim that awareness requires consciousness. These are cases in which a subject—generally the victim of a brain lesion—looks directly at an object but fails to experience the qualia usually associated with seeing that object, while nonetheless having his behavior in some way influenced by the presence of that object. This suggests that he is at some unconscious level aware of that object even though he has not seen it in any traditional sense. For example, a subject may insist that the paper in front of him is completely blank, but when forced to guess what is written on it he will guess correctly (at a rate well above random) that there is a large red $\times$ there. These cases have become darlings of the consciousness literature in recent years, as they both make it easy to isolate the notion of qualia and lend solid support to the idea of unconscious mental activity.[7] Some have used this phenomenon to argue that consciousness is not a necessary condition for awareness, but there is another explanation of blindsight available that lets us avoid this conclusion. If blindsight is interpreted as affecting the subject's dispositions to

[7]Especially helpful writings on blindsight include Lawrence Weiskrantz's fascinating and detailed accounts of blindsight in his patients [114, 113], and Charles Siewert's thorough analysis of the implications of blindsight for qualia-based consciousness [99, especially ch. 3].

behave in certain ways rather than affecting his awareness, then we are able to preserve the idea that awareness requires consciousness. I will explain what I mean. Certainly a subject with blindsight would have his neural firing patterns modified as a result of looking at the large red × even if he never experiences large, red, ×-like qualia (this is just to say that the image of the × has to enter his cognitive system somehow), but there is no need to think of this change in firing patterns as producing awareness. The firing patterns within his brain now dispose him to answer "there is a large, red X" when pushed to describe the seemingly blank paper, but this is no different than if he were primed to give this answer by hypnosis, by prior exposure to red × shapes that he actually sees (rather than merely "blindsees"), or by reading the words "there is a large red 'X' on the paper before you" displayed on a computer screen next to the paper. There is no overriding reason to think of the subject as being *aware* of the red × in any of these cases. Given this, and given the fact that blindsight patients rarely (if ever) spontaneously act on objects within the "blind" portion of their visual field, my claim that consciousness is necessary condition for awareness is not unduly threatened by the occurrence of blindsight.

Is consciousness a sufficient condition for awareness?

I don't think that consciousness is a sufficient condition for awareness. Certainly most cases of consciousness do have concomitant awareness: when tasting a strawberry I become conscious of a particular array of flavors and textures in my mouth, and I am presumably aware of the fact that there are certain flavors and textures on my palate. I might also be aware of the fact that I am specifically tasting a strawberry (though this need not be the case), and I might even be aware that I am aware of something. Considering the

possibility of meta-awareness, or awareness of awareness, there are probably a large number of levels of awareness that could come into play in this case, and indeed in any scenario involving qualia. But must awareness accompany consciousness in *every* case? It is difficult to come up with examples of consciousness without awareness, but I am not convinced that it is impossible. Think of a case in which someone is thoroughly shocked or stunned by what he sees or hears. These visual or auditory qualia register on the person, but if the input is strong enough and unexpected enough perhaps it could produce a sort of mental paralysis, in which no awareness occurs along with the qualia. Episodes in which thought processes halt completely are generally very short, but they do occasionally occur. Though awareness may follow soon after, it does seem that these would be cases in which there are brief moments of consciousness without awareness, as long as awareness is understood as involving only propositional thought. So it looks like consciousness is not a sufficient condition for awareness (though cases involving the former without the latter are extremely rare).

Let's pause for breath. So far I have argued that consciousness is a necessary but probably not sufficient condition for awareness. Or to look at the relata from a different perspective, awareness is a sufficient but probably not necessary condition for consciousness. This suggests that consciousness and awareness are not the same thing, despite frequent claims to the contrary in the literature. And if I am wrong about blindsight—if instances of blindsight do in fact produce real awareness without involving corresponding qualia—then this is even more evidence for the importance of distinguishing consciousness from awareness. That they are separate phenomena seems reasonable, given that they involve

very different sorts of things: saying that someone is conscious means he is experiencing at least one sight, sound, smell, or other quale, whereas saying that he is aware means that he knows (or thinks he knows) some fact about the world.

3.2.3 Consciousness is not introspection or higher-order states

A plurality if not a majority of analyses of consciousness focus on what are called "higher-order relations." The idea is a very simple and very old one—Güven Güzeldere [41] argues that its roots can be found in both Descartes and Locke,[8] while others trace it as far back as Aristotle. The general concept behind higher-order relation theories of consciousness is something like this. If a person has a mental state x, x is not a conscious state unless that person has another mental state y—a "higher-order" mental state—which is directed at x in some particular way. For example, if Alice has a mental state that represents the fact that her cat's fur is soft—and this state can be either propositional (i.e., she thinks "my cat's fur is soft" either in English or in her favorite dialect of mentalese) or qualitative (i.e., she simply feels the softness of her cat's fur)—the mental state will only be conscious if she has another mental state that somehow "observes" the first mental state. Different higher-order theorists describe the exact relation between the higher-order state and the lower-order state differently, but they all agree that mental state y has to be *directed at* or *represent* mental state x in some way, and that mental state y endows mental state x with state-consciousness.

[8] Güzeldere may have this passage from Locke in mind. I don't think it is completely unambiguous, but it could certainly be taken to suggest that he supports a higher-order view of some sort: "consciousness ... is inseparable from thinking, and, as it seems to me, essential to it: it being impossible for any one to perceive without perceiving that he does perceive. When we see, hear, smell, taste, feel, meditate, or will anything, we know that we do so." [59, p. 39]

These higher-order relations sound an awful lot like old-fashioned introspection, and higher-order definitions of consciousness can almost always be recast in terms of introspection: if you can introspect in such a way as to "observe" your own mental states, then those mental states are conscious. The state consciousness thereby produced is then sufficient to make you intransitively creature conscious. Given the virtual equivalence of these terms in this context, for ease of discussion I will use "introspection" and "higher-order theories" interchangeably.

What, then, can we say about introspection-based accounts of consciousness? In short, they don't work. Given both the strength of support that introspection-grounded analyses of consciousness have received in the literature,[9] and my own reliance on introspection as a tool for consciousness research, this claim may be surprising. But there is a very simple reason why they don't work. Introspection can be thought of as a special case of awareness: it is *internal* awareness rather than the more standardly used *external* awareness. That is, rather than making us aware of facts about the outside world, introspection brings to our awareness facts about our own mental states. The arguments I gave above concerning awareness in general also apply here, so introspection cannot be the same thing as consciousness.

But there are a number of other arguments against introspection or higher-order theories of consciousness available. Given the popularity of such theories, these arguments are worth examining in some detail. One such argument works especially well against the higher-order theory proposed by William Lycan. Lycan explains consciousness thus:

[9]See Peter Carruthers [14], Paul Churchland [18, 19], and William Lycan [60, 61] for just a few of the many arguments that have been made in favor of introspection-based definitions of consciousness. Block [9] and Dretske [29], on the other hand, are prominent critics.

> Consciousness is the functioning of internal attention mechanisms directed upon lower-order psychological states and events, [and] attention mechanisms are devices that have the job of relaying and/or coordinating information about ongoing psychological events and processes. [61, p. 1]

One might immediately wonder how the mere "relaying of information," even information related to internal processes, could be sufficient for producing consciousness. After all, the lower-level state is presumably relaying information as well, and on Lycan's theory it produces no consciousness by itself. Why would two relays succeed where one relay fails? I think we can safely assume that Lycan's general description of the phenomenal aspect of consciousness—his sense of what it's like for a system to be conscious—is similar to the one that I have proposed, but this makes his analysis of the *cause* of consciousness quite puzzling. It is just not clear how knowing something about a mental state could grant that state qualitative properties. Knowing something about an object doesn't give that object any qualitative properties (e.g., my knowing that my desk is made of maple doesn't suddenly give the desk qualitative properties), and it is not obvious why Lycan thinks that knowing something about a mental state should be any different than knowing something about a piece of furniture. And even if this *is* how mental states come to have qualitative properties, neither Lycan nor anyone else has given a satisfactory account of *how* such a bizarre causal connection would operate. Why would knowledge about a lower-level mental state cause that state become state conscious, and what sorts of knowledge about that state would be sufficient to invoke state-consciousness? Would I just have to know that I had it, or would I have to know when I had it, or whence it came, or how it related to other mental states? Lycan leaves these questions unanswered.[10]

[10]Curiously, Lycan does not shy away from—indeed, he embraces—a theory of mind that leaves these ques-

There is a stronger argument that can be used not only against Lycan's version of introspection, but also against all other versions I know of. The argument involves corrigibility of mental states. It is clear that our lower-order mental states can come unglued from the reality that they purport to represent. No philosophical problems are raised when we suffer visual hallucinations or hear sounds that are produced entirely by the imagination. Dreaming and daytime reverie provide plenty of cases in which our mental states do not accurately represent reality. Similarly, there is no difficulty with our thinking false propositional thoughts—that is, the possibility of unsatisfied intentional states poses no particular problem. But on introspective accounts of consciousness, the possibility of inaccurate *higher-order* mental states has troubling consequences. A lower-order state of pain is not conscious until we have a higher-order state that represents it or points to it. But what if we have a malfunctioning mental monitor that produces a higher-order state which points to a non-existent lower-order state? It looks as if you could be conscious of a state that you did not in fact have. You could have the qualia associated with a pain without having any mental state (such as a state that represents an injury or body part) that has that pain as a qualitative property. And that makes no sense. A person simply cannot *think* he is experiencing some qualia when he in fact is not. I think this argument alone is enough to discredit higher-order or introspective theories of consciousness.

Here is another argument derived from the appearance/reality distinction that

tions open. He promises that a fully-developed theory of representation will address all lingering worries we might have about the cause or nature of consciousness, but he gives no indication of how such an explanation would work or what it would look like: "[T]he mind has no special properties that are not exhausted by its representational properties. It would follow that once representation is eventually understood, then not only consciousness in our present sense but subjectivity, qualia, 'what it's like,' and every other aspect of the mental will be explicable in terms of representation, without the positing of any other ingredient not already well understood from the naturalistic point of view." [61, p. 2]

served as the crux of the previous argument. As already explained, introspection-based theories state that a subject can have a state-conscious mental state x by having a higher-order state y directed at the lower-order state x. But if there is no third-order state z directed at y, then the subject will not be conscious that x is a conscious state. Let me spell this point out for clarity. Because y has no higher-order state directed at it, y will not be state conscious. And if y is not state-conscious, then the subject cannot be aware that y exists. And then he cannot know that x is state-conscious. This point is important because it puts the subject in a peculiar position: he experiences the qualia associated with x but cannot know that x is a state-conscious state. That possibility simply runs contrary to our daily experience. Whenever some state has qualitative properties, we know that state to be state-conscious. There is never any doubt about this. When I taste chocolate, I not only have certain phenomenal experiences (i.e., the mental state produced by the chocolate has qualitative properties), but I also know that I am conscious of the taste of chocolate. I may not realize that it is chocolate that I am eating, but I would at least know that I am experiencing *some* flavor. And this introspection-based theories of consciousness would seem to disallow this last bit of knowledge—I would not my first-order mental state (the state with the distinctive chocolatey qualitative properties) was conscious. Although the problem I am pointing to is somewhat difficult to express, it will be obvious to anyone who considers any qualitative experiences they have had. Once understood, it becomes an obvious and devastating flaw. The possibility of there being no third-order mental state z suggests that there is a way something *actually seems* to me (the feel of my cat's fur actually seems soft, as I can tell by the "soft" quale that I feel when I stroke it), and a

way that it *seems to seem* to me (lacking any mental state z directed at mental state y, I am not aware that I have a higher-order mental state directed at the mental state x that represents my cat's fur being soft, and so I am not aware that the "soft" quale is part of any state-conscious state of mine). And whatever it might mean for there to be a distinction between the way that something *seems* to me and the way that thing *seems to seem* to me, any theory that allows such a distinction is suspect from the very beginning.

This last point should not be confused with the similar-looking but importantly different claim that qualia need not be produced by authentic perceptions. That is, qualia that arise during sleep or hallucination are just as legitimate and consciousness-sustaining as those produced by the perception of actual objects. This second claim is much more common than the point I am making with the preceding argument, and dates back at least as far as Descartes: "I am now seeing light, hearing a noise, feeling heat. But I am asleep, so all this is false. Yet I certainly *seem* to see, to hear, and to be warmed. This cannot be false; what is called 'having a sensory perception' is strictly just this." [27, p. 19] The point that my argument is based on is closer to the (true) claim that qualia reports are incorrigible, such as is suggested by Hilary Putnam:

> [O]ne can have a "pink elephant" hallucination, but one cannot have a "pain hallucination," or an "absence of pain hallucination," simply because any situation that a person cannot discriminate from a situation in which he himself has a pain *counts* as a situation in which he has a pain, whereas a situation that a person cannot distinguish from one in which a pink elephant is present does not necessarily *count* as the presence of a pink elephant. [80, p. 28]

Putnam's view seems to be that qualitative experiences cannot be false in the way that a belief or a perceptual experience can be. My closely related point is that when one experiences qualia, one cannot doubt that the state that has that qualitative property

is a conscious state. Just as a pain cannot be a hallucination, a pain also cannot be unexperienced or unfelt or hidden from consciousness. Unconscious phenomena do certainly exist, and they can probably leave the same emotional scars and produce the same emotional burdens as real, felt pains are capable of, but they are not *painful* in the same way. They are not examples of qualia.

Another way to show that certain introspection-based theories are wrong—or at least superfluous—is to show they reduce to qualia-based theories. This is the case with David Armstrong's theory of consciousness. Armstrong's theory and others like his are perfectly acceptable by my lights, but only because despite initial appearances, they ultimately rely on qualia. Let's look at Armstrong's theory, and an example he uses to support it, in more detail.

Armstrong [2] claims that we can have a minimal sort of consciousness when we have a simple lower-order state, but that we don't achieve full-fledged consciousness, or what he calls "introspective consciousness," until an appropriate higher-order state comes into play.[11] The lower-order state is enough to direct our activities to some degree, but it doesn't provide us with any real awareness of what is going either in our bodies or around us. His well-known example (which I have already discussed briefly in chapter two) involves a tired long-distance truck driver who drives without incident for hours without being aware of either his bodily movements, the decisions he makes while driving, or the passing scenery. While there is clearly some perception taking place—i.e., the driver possesses some of Armstrong's *minimal* consciousness—he lacks any trace of the more important and vivid

[11]He actually divides sub-introspective consciousness into *minimal* and *perceptual* forms, but because the distinction is not crucial for my purposes I will conflate the two.

introspective form of consciousness. Employing a term often used to describe states similar to this one, the sleepy driver is "zombified" while driving with minimal consciousness: he behaves in the ways we expect him to, but without any awareness of what he is doing. Eventually we would expect the driver to "come to," snapping back into full introspective consciousness.[12]

Armstrong uses this example to demonstrate the difference between minimal and introspective consciousness, and ultimately to support his view that higher-order states are a necessary prerequisite for full, introspective consciousness. But another lesson can be drawn from the example. The most prominent feature of Armstrong's introspective consciousness seems to be that it involves qualia; all that changes when the truck driver shifts from minimal to introspective consciousness is that his mental states are suddenly granted qualitative properties. This suggests that what is really important about consciousness—what allows us to immediately differentiate real, full-fledged consciousness from mere "zombie" consciousness—is the presence or absence of qualia. So Armstrong's example really seems to be drawn from two independent claims. First, he is claiming that full consciousness is produced when a higher-order state is directed at a lower-order state. I hope to have shown through my arguments above that this position is wrong. Second, he is pointing out that the important subjective difference between minimal and introspective consciousness—that is, the difference *to the person who shifts between these two states of consciousness*—is that the latter involves qualia and the former does not. And on this point, of course, Armstrong is exactly right. So we can occasionally salvage some helpful ideas from certain

[12] Armstrong carries this distinction into fascinating territory when he then speculates that introspective consciousness appeared later on the evolutionary scale than minimal consciousness, and therefore is relatively rare among animals with more primitive central nervous systems than ours [2, p. 60].

otherwise-problematic higher-order or introspective theories of consciousness.

There are further arguments against introspection-based consciousness for which I unfortunately cannot claim credit. Two common criticisms take the form of difficult questions that the supporters of higher-order theories have yet to address. First, how is it that a mental state can be affected simply by the fact that another mental state is pointing at it? Why would mental state x gain qualitative properties by virtue of it being the subject of higher-order mental state y? How is this causal relation supposed to work? Secondly, assuming that all mental states are grounded in or supervenient on physical states in some way (this criticism does not depend on any particular solution to the mind/body problem, as long as brain states enter into the equation *somewhere*) why is it that only certain physical states are capable of producing mental states with qualitative properties? Why doesn't monitoring of my digestive processes produce qualia associated with the amylase-induced breakdown of complex carbohydrates into simple sugars, the enzymatic breakdown of proteins into amino acids, or the absorption of vitamin C? These questions do not of course constitute a knock-down argument against introspection-based consciousness, but they do demonstrate some of the difficulties faced by the theory.

Perhaps the most common criticism of all is that higher-order theories lead to a regress problem. The problem runs like this. Higher-order theories claim that a mental state x is only state-conscious if it is monitored by some higher-order mental state y. But this kind of monitoring sounds like a form of perception, and all perception involves qualia. So y must have qualitative properties as well, and then the regress sets in. For y to have qualitative properties, y must in turn be monitored by an even higher-order state z which

would also have to have qualitative properties and hence would have to be monitored by another mental state w, and so on *ad infinitum*. I have presented this argument because it is so frequently used against introspection-based theories, but I have to admit to not being fully convinced by it. I think there might be room to argue that the monitoring involved in higher-order theories need not involve qualia (after all, wouldn't blindsight count as a form of qualia-free perception?) and so the regress might never get off the ground. It is actually somewhat difficult to think of cases in which we experience qualia associated with higher-order states. Think of a time when you realized that you were hungry. You certainly felt the hunger itself (a quale associated with a first-order state), but did you experience a quale associated with the *realization* that you were hungry? I'm not sure that we usually do. If there *were* a quale associated with this higher-order state of realizing you were hungry, it would presumably be different from the qualia produced by the realization that you were, say, tired or angry or cold. But it's not clear that we experience any qualia in these cases over and above the qualia associated with being hungry, tired, angry, or cold (i.e., first-order qualia). So I don't place much stock in this argument against higher-order theories. However, there is a variation on the argument that might be more effective.[13] Instead of relying on an infinite regress, the enemy of higher-order theories can simply point to the apparent absurdity of the claim that I cannot feel a conscious pain unless I also have an unconscious belief that I am having that pain. Regardless of whether this form of the argument works better than the infinite-regress version, I think that the other criticisms I have canvassed are sufficient to defeat introspection-based theories of consciousness.

As a final note, I want to make clear that I do not want to deny the existence of

[13]This variation has been used publicly by John Searle, though it may not have appeared in print.

higher-order states, or to claim that higher-order relations are completely unhelpful either in daily life or when studying the mind. On the contrary, they do exist and they serve as indispensable research tools in that they let their subject know something about his mental states. Do you feel out of sorts but don't quite know why? Introspect and maybe you'll realize that you are hungry as soon you form a second-order state directed at your first-order (but possibly state-unconscious, at least up until then) state of hunger. To reiterate a point I argued for in chapter one, introspection is a useful technique for finding out what mental states you happen to be in. It is not a trivial component of our repertoire of mental capacities; it may very well be something that distinguishes humans from less mentally sophisticated animals whose unilevel minds lack the ability to form higher-order mental states. So I think that higher-order theories do play an invaluable role in helping us to get a fuller understanding of the mind. They just don't contribute anything to our understanding of consciousness.

Chapter 4

Machine consciousness

> *How it is that anything so remarkable as a state of consciousness comes about as a result of irritating nervous tissue, is just as unaccountable as the appearance of Djin when Aladdin rubbed his lamp.*
>
> Thomas H. Huxley [48]

Huxley's astonishment notwithstanding, if we know anything about how the brain and mind work, we know that irritating nervous tissue absolutely does produce consciousness. Once this is understood, an obvious question presents itself: what other sorts of things can be irritated in such a way as to produce consciousness? Perhaps the most frequently suggested non-neuronal candidate for producing consciousness is silicon. In this chapter I will examine two sets of arguments against the prospect of a computer sustaining a conscious mind. In the first half of this chapter, I will defend John Searle's famous Chinese room argument against a set of criticisms leveled at it by Georges Rey. The Chinese room has faced as many attacks as any contemporary philosophical argument I can think of, and it would be impossible to address each of these individually. But Rey assaults the position from a variety of angles and uses a number of different approaches. I hope that by discussing

his points in some detail I will be able not only to give a sense of the arguments used by the anti-Searle camp, but also convey some general strategies that can effectively be used against these arguments. If I do my job properly, this defense of Searle should serve as a foundation upon which further responses can be built against any attack on the Chinese room—or at least any attack that I know of.

There are many ways of designing a computer, but most hardware designs fall under one of two high-level approaches to computer architecture. Generally speaking, computers either process data serially, following one instruction at a time (this strategy is also known as von Neumann architecture, after computer pioneer John von Neumann), or else they process data in parallel, using interconnected "nodes" that operate independently but whose output serves as input for other nodes. The Chinese room argument is often thought (incorrectly, as I will later explain) to address only the prospect of conscious minds in serial computers, so in the second half of this chapter I will present a series of arguments against the possibility of a human-like mind being produced by a parallel-processing computer.

4.1 The Chinese room argument

John Searle's Chinese room argument has become something of a lightning rod in the literature of artificial intelligence and philosophy of mind. It has attracted countless attacks by supporters of artificial intelligence while being celebrated by their opponents as irrefutable proof that the Strong AI project—a term that I will soon define—is doomed. My take on this seemingly endless debate is that Searle is mostly right and that arguments against his view are generally the result of a misunderstanding either of the definition of a

computer program or of what exactly the argument is meant to demonstrate. To support my claim, I will first describe the concepts of Strong and Weak artificial intelligence and review the mechanics of the Chinese room argument. I will then present three objections given by Georges Rey and show why each of the objections fails. Finally, I will turn very briefly to the troubling implications of Searle's argument, describing why it makes two of our most deeply held convictions appear to be incompatible with each other.

4.1.1 Strong and weak artificial intelligence

In order to understand what the Chinese room argument is intended to show, we must first review the notions of "Strong" and "Weak" artificial intelligence, or AI. These terms were originally coined by Searle but have since gained currency among the artificial intelligence and philosophy of mind communities. Weak AI consists in the relatively uncontroversial claim that computers can be useful tools for understanding the brain. Specifically, proponents of Weak AI claim that the mechanical workings of the brain can be simulated on a computer with a sufficient degree of accuracy that certain empirical data gathered from the computer model are likely to apply to the brain as well. For example, a computer could successfully be used to track blood flow in the frontal lobe or to model fetal development of the cerebral cortex, just as a computer can be used to track traffic patterns in Phoenix or to determine the strength and sway characteristics of a bridge before it is built. Searle takes no issue with weak AI, and, like most people involved in the broad area of cognitive science, considers it a valuable research tool.

Strong AI is much more fiercely debated. Searle defines it as the theory that the running of any computer program that is designed to accurately mimic the brain actually

produces a mind with conscious states: "the appropriately programmed computer really *is* a mind, in the sense that computers given the right programs can be literally said to *understand* and have other cognitive states." [93, p. 417] While a computer is needed to run any such program, the emphasis of Strong AI lies more on the program than the computer; a program that creates conscious states when running on a Macintosh would also create conscious states if adapted to run on a IBM mainframe or on my Palm Pilot. However, the same Macintosh, IBM, or Palm Pilot, would *not* be conscious while running a word processor, a flight simulator, or any other non-brain-modeling program. While the supporter of Strong AI clearly cannot attribute consciousness to the program itself (for the program is just an abstract entity—a mere list of commands), he does attribute consciousness to the appropriate program running on any compatible hardware. To put the point a different way: Strong AI claims that the program as scribbled out on paper, read aloud by a programmer, or even entered as static data into a computer is not conscious, but that as soon as the program is run, as soon as its commands are executed by a digital computer of any kind, consciousness appears.

One element of Strong AI that many find unclear, but which proponents of the theory seem reluctant to dedicate much discussion to, is the question of what exactly emerges when a computer runs a program of this sort. Philosophers of mind are careful to keep terms such as "consciousness," "intentionality," "understanding," and "mental states" quite separate (even if, as sometimes seems the case, every philosopher of mind defines them slightly differently), while artificial intelligence researchers tend to conflate them. This conflation may be deliberate, but I suspect it stems more from lack of careful thought about

what the terms really mean. In any case, Strong AI is of relevance to this discussion because consciousness is almost always included within the constellation of mental phenomena that is supposed to arise from the right sort of program. It is important to keep in mind, however, that the Chinese room argument is directed at more than just the possibility of computer consciousness—it also denies the possibility of a computer possessing real intentional states or semantic content simply in virtue of running a program.

4.1.2 The structure of the argument

Let us begin discussion of the now-legendary Chinese room argument by stating flat-out what it is meant to show: if the argument is right, then no computer that lacks the causal powers of the human brain can be conscious, no matter what program it runs. Unless explicitly described otherwise, all computers henceforth referred to should be assumed to *lack* these causal powers. This emphasis on causal powers is absolutely crucial to the argument, for Searle is quick to admit that any system, whether based in carbon, silicon, or any other building block, would enjoy a full mental life—consciousness and all—if it has the same causal powers as the brain. Oddly, this qualification is often missed or ignored by critics, and this mistake seems to be responsible for a large number of the attacks on the Chinese room.

A secondary thesis is that because no such computer can be conscious, studying computer programs will give us no insight into how human consciousness is actually produced, nor will it provide an accurate way to study human thought processes or cognition generally. Or, in Searle's own words: "Any attempt literally to create intentionality artificially (Strong AI) could not succeed just by designing programs but would have to duplicate

the causal powers of the human brain.... Strong AI has little to tell us about thinking, since it is not about machines but about programs, and no program by itself is sufficient for thinking." [93, p. 417] Now with these two conclusions in mind, let's look at the structure of the argument.

Searle picks a particular computer program—Roger Schank and Robert Abelson's model for reading and correctly answering questions about simple stories in English [92]—and sets up the Chinese room to disprove directly any claim that this model achieves real understanding. But because the conclusions drawn by the Chinese room argument are meant to apply to *any* computer program and not just Schank and Abelson's program, the version of the argument that I will present here will be simplified somewhat to make the broad scope of the conclusions more apparent.

With this caveat in mind, the main features of Searle's argument are these. Imagine Jones, a tweedy Brit who speaks only English, locked inside a windowless room. Imagine further that people outside the room, who don't know that Jones is inside, "talk" to the room by pushing through a slot in one wall slips of paper that have Chinese characters (call them "input characters") written on them. Inside the room with Jones is a book, written in English, that contains instructions which assign a second set of Chinese characters (call them "output characters") to every possible set of input characters. Upon receiving a slip of paper with input characters, Jones uses the instructions in the book to write an appropriate set of output characters on another slip of paper and shoves this response back through the slot in the wall. Unbeknownst to Jones, the input characters he receives are all well-formed and coherent questions in Chinese, and the characters he is instructed to provide as output

are equally well-formed, coherent, and most importantly, reasonable, answers in Chinese to those input questions. If asked "How are you?" he might unknowingly answer "Fine, thank you." "Seen any good movies lately?" might elicit "No, I don't get out much." The Chinese equivalent of nonsense words might result in a response asking for the question to be rephrased, or a suggestion that the questioner lay off the rice wine. The important thing to notice about Searle's argument is that Jones does not understand any of the characters he either receives or issues in response, and he need not even be aware that they are sentences in any language at all; for all he knows, they are random scribbles. To him, neither the input nor the output has any meaning. Despite this utter lack of understanding, Jones becomes so facile at this task of looking up meaningless (to him) scribbles in a book and writing out different but equally meaningless (to him) scribbles, he is eventually able to convince native Chinese speakers outside the room that the room itself can both understand and produce written Chinese. The room appears to have intentional states. It seems to know what it is being asked and it looks like it knows how to respond. The room appears to be intelligent, and to be able to think, and to be conscious.

Searle claims that all of these appearances are illusory. We have stipulated that Jones doesn't understand Chinese, and since even the strongest supporter of artificial intelligence would agree that neither the walls nor the book nor the air inside the room, considered apart from Jones, understand Chinese, the room as a whole must not understand Chinese. Furthermore—and this is the central point of the Chinese room argument—since the act of Jones following instructions in the book is analogous to the act of a computer following instructions in a program, Searle reasons that no computer can be conscious merely by

virtue of running a particular program. It is important to remember the earlier claim that a computer with the right causal powers *could* be conscious, and to contrast it with the point Searle is making here. The primary aim of this argument is to show that unless the physical apparatus that is the computer has causal powers capable of producing consciousness absent any program, the mere act of having it run a program, no matter how sophisticated, cannot produce consciousness within the computer. This is the first of the two conclusions that were laid out above. The second conclusion falls out of the first: because no program can produce consciousness in any computer it runs on, the study of programs cannot be of any help to us in understanding how the human mind works. The program model of conscious thought has no explanatory power:

> [W]hat is suggested though certainly not demonstrated by the example is that the computer program is simply irrelevant to my understanding of the story. In the Chinese case I have everything that artificial intelligence can put into me by way of a program, and I understand nothing; in the English case I understand everything, and there is so far no reason at all to suppose that my understanding has anything to do with computer programs. [93, p. 418]

Searle takes pains to point out that although he cannot prove that the Chinese room scenario sheds no light on how we understand language or partake in conscious thought, the simple fact that there is no real understanding of Chinese going on in Jones suggests very strongly that no model that deals only with the formal (i.e., syntactical) properties of words or objects can produce understanding.

A final note on the example: Searle is careful to clarify what he means by "understanding" when he says that the Chinese room doesn't understand Chinese. He admits that understanding is not a simple two-place predicate; there is plenty of wiggle-room when it comes to understanding. We can reasonably say that a person partially understands some-

thing, or mostly understands something, or slightly understands something. Moreover, what one person considers understanding may fail to satisfy another person's conditions for understanding. But Searle responds that there are certain cases where it is very clear that no understanding takes place. Neither his car nor his adding machine understands a thing, and the Chinese room is exactly the sort of case in which zero understanding occurs. With the intended conclusion of the Chinese room argument clear, let us turn now to Rey's complaints with the argument.

4.1.3 Criticism one: the Turing test is inadequate

Rey begins his critique by claiming that Searle misunderstands the claims made by supporters of Strong AI: "Searle's use of the example [of Abelson and Schank's research] actually flies in the face of what reasonable defenders of that thesis [i.e., the thesis of Strong AI] would explicitly claim." [84, p. 170] Specifically, he thinks that Searle attributes too much faith in the Turing test to adherents of Strong AI, where the Turing test is, in a nutshell, the test of whether a computer communicating by teletype can convince a person on the other end of the teletype that the computer is also a person.[1] That is, Searle uses the failure to find consciousness in a system that would pass a Turing test as evidence against Strong AI, whereas proponents of Strong AI consider the Turing test an insufficient means of testing for consciousness. To put Rey's claim even more simply: Strong AI has no quarrel with Searle if his project is simply to deny the consciousness of programs that pass the Turing test.

[1] For a full presentation of the structure of and criteria used by the Turing test, see Alan Turing's classic article [108].

According to Rey, the Turing test fails to establish consciousness under Strong AI because it is a test of behavioral properties whereas Strong AI is a thesis about functional properties. Behavioral theories are not concerned with *how* input (in the case of the Turing test, questions typed by a human judge) is transformed into output (the computer's response to those queries). They are interested only in *which* output a system produces given a particular input. Functional theories such as Strong AI, on the other hand, are concerned exclusively with *how* this transformation takes place. The Turing test alone would not be enough to convince a supporter of Strong AI that a machine is conscious or understands questions, for the machine might not be computing answers to those questions (i.e., transforming input into output) in quite the right way to manifest consciousness; despite giving convincing answers, it might not be running the right program to produce consciousness. Rey calls the thesis that the right type of program must be running in order for a machine to be conscious "AI-Functionalism," and claims that Searle seems not to have realized that Strong AI subscribes fully to AI-Functionalism. Strong AI would then grant that although the room could indeed pass a Turing test (for it is behaviorally equivalent to a conscious Chinese speaker), because it is probably running the wrong program—and this presumably means that Jones lacks the proper instruction book—it is most likely not conscious.

My first response to Rey's objection is to note that not all supporters of Strong AI place the bar for consciousness so high. A great number do seem to consider the Turing test to be the final arbiter of consciousness. Even Roger Penrose, who has argued *against* Strong AI in two books [76, 77], admits that the Turing test would be sufficient proof for him. However, there is an element of truth to Rey's remark in that there seems to be no

consensus of what *exactly* the Turing test would prove were a machine to satisfy it. Many philosophers and cognitive scientists feel that it is a good test of intelligence, but not of consciousness or intentionality (assuming that position to be coherent). Others claim that it does not even firmly establish the presence of intelligence. Alan Turing himself claimed somewhat cryptically that the question of whether a machine could think was meaningless and unanswerable.

In any case, even if Rey were right in thinking that most supporters of Strong AI demand more than success at the Turing test in order to grant that a machine is conscious, this is not a difficult attack to parry. To meet the objection that the Chinese room exhibits the behavior but not the functional structure of a conscious Chinese speaker, we can easily modify the argument to include whatever functional characteristics Rey or any other AI-Functionalist desires. Replace Jones's instruction book with one more similar to the program of the mind—whatever that might be—and Rey will find that Jones's knowledge of Chinese increases not one whit. If he objects that the whole von Neumann machine model is wrong and that the mind runs a program based instead on connectionist principles, this can be accommodated as well: give Jones a book that contains instructions for emulating a connectionist program on serial architecture, and have him "run" that program instead. Clearly we cannot execute a connectionist model directly, for that would require multiple copies of Jones within the room, corresponding to the large number of nodes that make up a connectionist machine. However, because we know that any connectionist program can be implemented on a standard serial computer, we also know that Jones can emulate any connectionist program Rey might have in mind.[2] The fact remains that no matter what

[2]This is an implication of the widely accepted though not formally proven Church-Turing thesis, which

English instructions Jones follows, he will never understand a word of Chinese.[3]

One might also respond to Rey's first criticism by asking him to clarify the claim, central to AI-Functionalism, that a program must be of a particular type in order to produce consciousness. In virtue of what exactly is a man-made program of the same type as the alleged program that produces consciousness by running in the mind? Certainly not in virtue of its overall input and output patterns, for Rey has already rejected behavioral tests for consciousness. Perhaps the program must be composed of particular algorithms, or maybe it must be line-for-line identical with the brain's mind-program. If Rey demands identity at too low a level of abstraction (e.g., if he specifies the individual commands that make up a consciousness-sustaining program), then he will find that he has limited the range of hardware on which the program can run. In other words, if a program must consist of particular commands in a particular order then it can run only on architecture that supports those commands, for not all computers recognize the same software instructions. Because the whole point of Strong AI is to strip away hardware or substrate considerations from consciousness, any significant limitation of the range of physical material that can sustain consciousness is a serious threat to the theory. This problem is mitigated somewhat by the fact that most programming languages can be emulated on most computer platforms. For example, perhaps we could make a Hewlett-Packard workstation conscious by emulating the programming language of the mind accurately enough so

states that any algorithm can be computed on a Universal Turing machine. Both von Neumann and connectionist computers are mathematically equivalent to Turing machines, so any algorithm that can be implemented on one architecture can also be implemented on the other architecture (though perhaps significantly more slowly).

[3]It is interesting to note the parallel between this objection of Rey's and the "brain simulator" reply that Searle responds to in "Minds, Brains and Programs." While the brain simulator reply demands that Jones implement particular *hardware* in order to produce consciousness, Rey demands that Jones implement particular *software*. In neither case is Jones's comprehension of Chinese affected.

that a program that is command-for-command equivalent to the mind's program would run smoothly on the Hewlett-Packard. The most obvious problem with this approach is that emulation significantly reduces the speed at which a program can be executed. It appears that the AI-Functionalist would either have to sharply limit the range of platforms on which a consciousness-producing program would be able to run (thereby rendering emulation unnecessary), or else produce a convincing argument for why speed of execution is irrelevant to the production of consciousness. The former is an unappealing option, for it would mean that seemingly arbitrary distinctions would have to be made between those substances that could sustain consciousness and those that could not. Is the AI-Functionalist really willing to say that any hardware that has six internal registers can be conscious, while those with only five cannot? Could he plausibly maintain that hardware that supports floating-point precision to twelve decimal places can be conscious, while hardware with precision to only eleven decimal places just misses consciousness, despite being behaviorally all-but-identical to the machine with twelve-place precision? The alternative option of explaining why consciousness can be present regardless of execution speed may be possible, but seems extremely difficult from an intuitive standpoint. It is hard for us to imagine having quite the clarity of consciousness that we currently enjoy if our brains were slowed to 20% of their normal speed, a reasonable factor by which a program might be expected to slow after being ported to a hardware platform that does not natively support the language in which the program is written. We have no idea whether our consciousness would continue to exist at all under such circumstances, but it seems overwhelmingly likely that any consciousness that was still present would be of a radically different character from that which

we presently experience. And that is enough to defeat the AI-Functionalist, for his claim is not just that the right program creates consciousness, but that it creates consciousness that is quale-for-quale identical to our own, given the same input.

Considering these difficulties, Rey's best bet is to maintain that programs are type-identical in virtue of implementing the same general algorithms. This is far from a precise requirement, as it is not at all clear at what level of abstraction the algorithms would have to be identical. This issue presents Rey with two problems that correspond neatly to the two general concerns Searle has with the Strong AI project. First, it seems unlikely that a principled reason could be given for why algorithms must be identical at a particular level rather than at any other level in order for two programs to be type-identical and therefore to produce the same sort of consciousness. This is a concern about whether minds really are programs at all. Second, how could the AI-Functionalist determine at which level of abstraction algorithms need to be identical, when other programs that are identical to the human mind-program at other levels of algorithmic abstraction are nonetheless behaviorally equivalent. This is a concern about whether the study of computer programs can be at all helpful in understanding the operation of the mind.

Let us examine each of these concerns more fully in order to make sure they are clear. Just as physical systems can be described at many different levels of abstraction (we can describe a car at the quark level, the molecular level, the spark plug/fan belt/fuel injector level, or the engine/transmission/chassis level, to pick just a few possible levels), so can computer programs be described at different levels of abstraction. A program that prints mailing lists could be described as just that, or as a program that reads names and

addresses off the hard disk, sorts them by zip code, puts them all in the Garamond font, and prints them out at two columns per page. Or we could move to an even deeper level by describing it as a program that reads straight ASCII files from any SCSI device, strips off file header information and resource forks, sorts the data with a bubble-sort/quick-sort hybrid, and so on. Now the AI-Functionalist thesis is that a man-made mind-program can only be conscious if it is type-identical to a real human mind-program, and specifically if it is type-identical in virtue of sharing algorithms. My first concern with this is that it is not clear that any principled reason could be given for why type-identity at a particular algorithmic level produces consciousness while type-identity at higher algorithmic levels does not. It is perfectly principled to claim that programs must be identical at either their highest level of abstraction (i.e., that they are merely behaviorally identical) or their lowest level of abstraction (i.e., that they are command-for-command equivalent), but the idea that their type-identity depends on identity at some seemingly arbitrary intermediate level of abstraction seems odd. This is certainly not meant to be an airtight argument against this possibility; it is just an explanation of why this approach does not seem, at least intuitively, terribly sound. The second concern is that it would seem to be an impossible task for the AI-Functionalist to determine exactly what the crucial level of algorithms is. Lacking any certain test for consciousness, the AI-Functionalist is likely to fall back on some form or another of the good old Turing test. But as both Rey and Searle emphasize, the Turing test is strictly a behavioral test, and as such lacks the power to discriminate between those man-made mind-programs that have the right sorts of algorithms and those that don't. Any program that is type-identical to the program in the human mind at the

right algorithmic level to produce consciousness will be behaviorally identical to countless other programs that are type-identical to the human mind-program only at higher levels of abstraction, and thus are not conscious. After all, two mailing list programs can be behaviorally identical even though one uses the bubble-sort mechanism for its sorting algorithm and the other uses quick-sort. So even if minds really are programs, and even if man-made programs can be conscious if they are type-identical to the program of the human mind at the right algorithmic level, the AI-Functionalist would be unable to determine exactly which man-made programs satisfy the Strong AI project, and so would not know either which programs are capable of sustaining a mind or which programs would serve as useful tools for modeling and understanding the mind.

4.1.4 Criticism two: language comprehension is not autonomous

Rey claims that the program run by the Chinese room is too impoverished to have any chance of achieving real consciousness. He feels that it would need to operate synergistically with other programs that run concurrently—programs for perception, belief fixation, problem solving, preference ordering, decision making, etc.—if were to stand a chance of producing consciousness. These other programs would run independently but would share their results with Jones's original language-processing program. Those other programs, Rey claims, are all related to and necessary for language use and comprehension:

> [W]e need to imagine [Jones] following rules that relate Chinese characters not only to one another, but also to the inputs and outputs of the other programs the AI-Functionalist posits to account for the other mental processes of a normal Chinese speaker.... [U]nderstanding a language involves being able to relate the symbols of the language to at least some perceptions, beliefs, desires, and dispositions to behave. [84, p. 172]

Rey also suggests that the understanding of language sometimes requires non-verbal responses (think of a request to pass the fried rice and snow peas). His point is that a Chinese room that runs only a language-processing program cannot be conscious because conscious language comprehension is not an autonomous process: "Searle's example burdens AI-functionalism with a quite extreme view about the 'autonomy' of language, a view that would allow that understanding a language need involve *only intra*-linguistic symbol manipulations." [84, p. 171]

This much of Rey's criticism is easily refuted with an adaptation of Searle's response to a similar criticism which he calls the "Robot Reply." Let Rey give Jones any number of books, each corresponding to a program distinct from the original language-processing program. By performing a single step from one book, marking his place in that book, performing a single step in the next book, marking his place in that book, and continuing to work his way through all of the books in this manner, Jones can, in effect, run many programs simultaneously. This technique of cycling through all of the programs, processing only a small portion of each before moving to the next and then returning to the first program at the end of each cycle, is known in computer science parlance as "multitasking," and allows any modern computer to run several programs simultaneously. Jones can then use the output from one book (say, the book representing the perception program) as input for another book (perhaps representing the language-comprehension program). By adding enough books we could give the Chinese room the full processing power of the human mind, abandoning completely any requirement that language processing be autonomous. But even after enriching Jones's program in this way, it seems very clear that he would still be utterly

in the dark. No matter how many books we give him, he will never understand Chinese.

I should take a moment to clarify an often-misunderstood element of this response. Consider a visual perception program that we might add to Jones's mix. This program would not receive actual images as input. Rather, it would receive incomprehensible (to Jones) descriptions or representations of visual scenes. These descriptions or representations could be, but need not be, in Chinese; they could just as well be any other non-English description. We cannot allow Jones to see actual images because this would allow the system to capitalize on intentionality that is *already present* within it—namely, the intentional state present in Jones when he perceives actual images. If Jones were given an instruction in the visual perception program book that said "if you receive the symbol ⊕ as input, give squiggle squoggle as output," he would likely come to understand that squiggle squoggle means "circle with a cross inside." An instruction containing the symbol ⊕ is no better than the instruction "if you receive a circle with a cross inside as input, give squiggle squoggle as output." The first instruction makes use of preëxisting intentionality, and the second draws on preëxisting semantic knowledge. A computer programmer trying to develop a conscious, comprehending program enjoys no such luxury. He cannot assume that there is any intentionality or semantic knowledge already present in the computer onto which he can latch additional intentionality or semantic knowledge. This maneuver is simply not open to him since his project is to introduce consciousness and intentionality into a system that originally has neither. The instructions for Rey's proposed visual perception program, then, must use only Chinese or other incomprehensible (to Jones) representations of visual patterns as input, rather than using the visual patterns themselves. This is not

as unreasonable a restriction as it may at first appear. After all, a similar translation from image to representation occurs when the human brain perceives visual images. The brain does not perceive objects directly; the retina creates neuronal impulses that represent objects, and it is these impulses rather than the objects themselves that serve as input for the visual centers of the brain. Of course, this is not to suggest that homunculi are needed either to perform the original translation from object to representation or to read the neuronal representation and translate it back into the image that is consciously perceived. I am simply pointing out that the brain does not operate directly on physical objects, but rather on neuronal impulses flowing from the retina.

So far, Rey seems to have served up an easily refutable argument. He then weakens his criticism further by committing a truly crucial error. He claims that proponents of Strong AI would demand that the collection of concurrently running programs in the Chinese room would have to include a program that provides *semantics* for the Chinese characters that Jones is wrestling with:

> Some people would even claim that you'd better have recursive rules of the form "F is true if and only if p" for Chinese F and (assuming that the rules are in English) English p: e.g. "'Squiggle squoggle' is true iff snow is white, 'Squiggle squoggle square' is true iff snow is not white." Indeed, I don't see why we oughtn't let the AI-Functionalist lay her semantic cards right out, as it were, on the table, and include recursive rules of the form "F *means* that p." [84, p. 172]

Considering that one of the main theses of Searle's argument is that syntax is insufficient for semantics, this attempt to introduce semantics into a system that is defined purely in terms of syntax is more than a little baffling. This argument is so bad that if it hadn't been presented unambiguously in the passage quoted above, I would be strongly tempted to think that I was misunderstanding Rey. Alas, that seems not to be the case. It

is quite clear that Rey wants first to explain to Jones what the Chinese symbols mean, and then to object that the Chinese room fails to do any damage to AI-Functionalism because Jones now understands what Chinese symbols mean!

This is a simple point, but it is so devastating to Rey's objection that I will state it once more: Rey wants to link Chinese semantics to English semantics—he wants to teach Jones what a Chinese symbol means by equating it with an English word or phrase. I grant that this ploy will indeed allow Jones to understand Chinese, but it is not a useful modification to Jones's program. The reason is essentially the same as the explanation for why the visual perception program cannot take raw images as input: Rey's proposal relies on semantics already present in the system—semantics that are not originally present in any computer on which a supposed "mind-program" could run. Since the Chinese room is described in such a way as to involve only manipulation of objects according to their *syntactical* properties, and since the issue of whether computer programs are able to recognize and make use of *semantics* is the issue in question, it makes no sense for Rey or AI-Functionalists to demand that Jones must have access to semantic as well as syntactic rules. This bridges precisely the gap between syntax and semantics that is in question; common sense tells us that a computer program involves only syntax,[4] and the burden of proof is then on the AI-Functionalist to show that semantics can be drawn from syntax. Rey needs to give some account of how semantics (i.e., the meaning of p that the F would take on as its own meaning, were Rey's "F means p" rules given to Jones) enter the computer in the first place.

[4]Searle has an interesting argument that even this is granting too much to the program. Because syntax is an observer-relative feature, a computer without any observers to assign syntax to its physical states has neither semantics *nor* syntax [95, ch. 9]. For space reasons, I will not discuss or evaluate this intriguing argument here.

4.1.5 Criticism three: the fallacy of division

Rey's final complaint is that Searle confuses Jones's properties with properties of the entire system: that Jones fails to understand Chinese is no evidence that the entire system of Jones, the book, and the walls, floor, and ceiling that make up the room fails to understand Chinese. Rey calls this the "fallacy of division." He does admit that Jones plays a central role in the operation of the room, analogous to the role played by the central processing unit (CPU) in a computer. Just as a computer's CPU serves to coordinate and control the computation of data received, stored, and processed (at least partially) by other subsystems within the computer, so Jones coordinates and controls the internal operation of the Chinese room. Rey calls this "the important role of deciding what finally gets said and done, when." [84, p. 174] However, he explicitly denies that consciousness is a property that could *only* be found in Jones or a CPU. Rather, consciousness is a property that could be held by the entire Chinese room system or by an entire computer system: "The AI-Functionalist no more needs or ought to ascribe any understanding of Chinese to this latter part of the entire system [i.e., to Jones] than we need or ought to ascribe the properties of the entire British Empire to Queen Victoria." [84, p. 174]

Rey then supports his point with a cryptic reference to a homunculus argument advanced by Daniel Dennett: "were the CPU *required* to have the mental states that are ascribed to the system as a whole, we would seem to be confronted with the familiar regress of homunculi that, as Dennett has pointed out, AI-Functionalism is so expressly designed to avoid."[84, p. 174] As Rey does not spell out this concern about homunculi in any detail, it is not immediately obvious what he is driving at. Perhaps his worry is this: if we require

a subsystem within the system to be conscious in order for the system to be conscious, then it appears that we would also require that a smaller subsystem *of the first subsystem* be conscious, if the higher-level subsystem is to be conscious. In other words, if we claim that Jones must be conscious in order for the room of which he is a subsystem to be conscious, then perhaps we must also claim that some subsystem of Jones must be conscious in order for him to be conscious. A regress might form in this way, with smaller and smaller subsystems or homunculi needing to be conscious in order for the overall system to be conscious. Of course, this problem would only arise if each subsystem must be conscious *to the same degree* as is the system one level up. This is important because Dennett's solution to this regress is to have each subsystem be dumber, or somehow less fully conscious, than its parent system. This allows him to consider the lowest level of subsystems to be completely mechanistic and utterly unconscious, thus ending any regress (or perhaps more accurately, preventing the regress from ever getting started by discharging homunculi completely from the system).

The homunculus concern aside, Rey's fallacy of division argument seems similar to a common attack on the Chinese room argument that Searle has dubbed the "Systems Reply." Searle's response to this reply works well against both Rey's fallacy of division and my interpretation of his version of Dennett's homunculus problem. Searle suggests that we eliminate the division by having Jones memorize the rules of the book and then continue to process Chinese characters as before, but now from memory. The result is that even when the fallacy of division is eliminated, comprehension is still missing. Some may object that Jones has only internalized a portion of the room—what about the walls, the

air, the floor? It is hard to know how to respond to this other than by saying that it seems inconceivable that these elements of the room contribute in any way to any consciousness the system might enjoy. The book and Jones are the only components of the system that could possibly be relevant—the room is there just to provide a screen that prevents the native Chinese speakers outside from knowing what is going on inside. Jones could just as well be hiding behind a big rock, or communicating by Chinese teletype from another city, or, as in the Robot Reply, receiving input from a television camera and sending output through a remote-controlled Etch-a-Sketch. We certainly wouldn't suppose that a rock that played this role would contribute to the consciousness of the system, and we wouldn't suppose that the teletype equipment, television camera, or etch-a-sketch would add to consciousness in any way. When Jones internalizes all relevant portions of the system yet still finds himself ignorant of the semantics of Chinese, the fallacy of division is proven not to be a fallacy at all. Certainly Rey is right in denying that all properties of the British Empire must apply to Queen Victoria, but there seems every reason to suppose that Jones must understand Chinese in order for the system to understand Chinese.

Rey anticipates this response to the fallacy of division objection, and claims that it fails, for were Jones to internalize *all* relevant elements of the system, he would have to have memorized not only the rules for correlating Chinese symbols, but also the other programs for perception, belief fixation, and semantic meaning of Chinese that Rey requested in his autonomy of language criticism. And once Jones incorporates all of these additional programs, he certainly has all of the linguistic and cognitive capacities of a native Chinese speaker:

> There we have our flesh and blood person, with as fully a biological

> brain as Searle would demand; he's now not merely following rules [by] *reading* an external manual, but has, by memorizing them, *internalized* them. And remember: they are not the mere "correlation" rules of the sort Searle provides, but full recursive grammatical and semantical rules of the sort that many philosophers of language would demand. The person in the room has done all this and *still* he doesn't understand Chinese? What more could Searle conceivably want? Surely this sort of internalization is precisely what's *ordinarily* involved in understanding Chinese. [84, p. 174]

The response to Rey's claim is obvious. We can certainly have Jones internalize programs for perception, belief fixation, preference ordering, and decision making, but since we have already rejected as inappropriate any program that would give Jones semantic information, there would be much that Jones lacks relative to a native Chinese speaker. The resulting deficiency of semantic content is absolutely critical, for knowledge of semantics is exactly what is at issue here. So I claim that that Searle's response to the Systems Reply is adequate to counter the fallacy of division argument.

If I understand the argument of Dennett's that Rey invokes (and Rey spends so little time developing it that I can't be sure that I do), then Searle's response to the Systems Reply handles it as well. It does so by pointing out that the regress of subsystems that need to be conscious stops at Jones. This is because Jones can internalize any element of the Chinese room that Rey or Dennett think relevant to the production of consciousness, whereas no subsystem of Jones would be able to internalize, simulate, or otherwise absorb all elements of *him* that are relevant to the production of consciousness. It seems quite obvious that his (living, properly functioning) brain is the only component within Jones that is necessary for consciousness, and there is no subsystem of Jones that is capable of internalizing the functioning of his brain other than his brain itself. Or to cast this point

in Dennett's terminology, Jones himself is the lowest level homunculus within the Chinese room system.

4.1.6 Implications of the Chinese room argument

The Chinese room argument has drawn an enormous amount of criticism in the twenty years since it first appeared, but I think much of that criticism derives from a misunderstanding of what exactly the argument is supposed to show. Because of the pervasiveness of this mistake, I will end my discussion of the Chinese room by repeating once more the intended conclusion of the argument and contrasting that conclusion with some points that the argument was *not* intended to demonstrate.

In its original form, the Chinese room was designed to make a very simple point: *systems cannot have intentional states merely by virtue of manipulating symbols according to their syntactic properties.* Or to put it epigrammatically, *you can't get semantics from syntax alone.* I have extended the argument (in a manner that Searle has endorsed) so that it shows that syntactic manipulation is not only insufficient for the production of *intentionality*, but is also insufficient for the production of *consciousness.* But that's it—that's the extent of the intended conclusion of the argument.

The most common way that the argument has been misunderstood over the years is that it has been mistakenly thought to be aimed at showing that *computers cannot, under any circumstances, have intentionality or consciousness* (in the rest of this discussion I will use "consciousness" to stand in for "intentionality or consciousness"). The distinction between the intended conclusion and this mistaken conclusion seems to have evaded many critics of the argument, but there is in fact a huge difference between the two claims. The

real conclusion just says that symbolic manipulation is not sufficient for consciousness, so we cannot say that a computer or any other physical system has consciousness simply in virtue of fact that it is running a particular program. There may in fact be consciousness in all sorts of physical systems, but any system that has consciousness will have this feature only because of the *causal powers* of the physical components of that system. To repeat: no amount of symbolic manipulation can produce consciousness, but the right causal powers can. Part of the mistake that Searle's critics make is that they fail to see a very important implication of this conclusion, namely *computers may very well be able to be conscious.* But if a computer *were* conscious, it wouldn't be so in virtue of being a computer or of running a particular program. It would be so because the physical components and the arrangement and behavior of those components give the computer the right causal powers to produce consciousness.

Here is another way to see this point. We know that different computers can run the same program, and the same computer can run different programs. This flexibility (together with the almost incomprehensible speed with which they perform the mathematical and logical operations that ultimately make up all programs) is what gives modern computers their astounding power. But now consider: two programs can run identically (except for likely differences in speed) on two computers that are made up of vastly different components. The same word processing program may look and act exactly the same on your computer as it does on mine, even though you have a 200MHz NuBus-based Motorola PowerPC CPU with 16Kb of level one cache and 3.3 volt EDO RAM, and I have a 300MHz Intel Celeron CPU operating on a PCI bus with 32Kb of cache and PC133 RAM. This simi-

larity at the user-level is possible because the program imposes in a certain sense the same functional organization on the lower-level hardware. For any given piece of software, this functional organization can generally be applied to large variety of systems, as is demonstrated when a particular program is "ported" from one hardware platform to another. But what the Chinese room shows is that this capacity for changing a system's functional organization—powerful though the capacity may be—is not sufficient to produce consciousness. A computer that is running a particular program cannot be made conscious purely by the fact that it is functioning so as to manipulate symbols in a particular way according to their syntactic properties. That is not to say that the computer cannot be conscious in virtue of certain other properties it may have, but its functional organization alone will not be enough. Any consciousness it may enjoy is brought about not by the software that it is running, but rather by the causal powers inherent in its hardware.

I hope this discussion has cleared up a common confusion about the Chinese room. Searle does *not* in fact use the Chinese room to argue that computers or machines in general cannot in principle be conscious. On the contrary, he has said explicitly that with the right causal powers, virtually any system can be conscious. And on this point he is quite obviously right; considering that the brain is nothing more than a physical system, and that it is fully capable of producing consciousness, it is hard to imagine how one could take seriously the position that computers or machines could not even in principle be conscious.

Understanding this mistaken criticism of the Chinese room can help to resolve an apparent contradiction that people often run across when thinking about the issue of computer consciousness. The contradiction is this. On the one hand, we firmly believe

(at least, most of us do) that consciousness is produced entirely by a purely mechanistic brain. The mysterious behavior of physical particles at the quantum level and lingering thoughts about dualism aside, the brain seems to act in a strictly mechanical way—a way that seems in principle (if not in practice, given the current state of science) to be completely understandable and predictable. There is nothing spooky, ethereal, or ineffable inside our brains that gives us consciousness; the task is accomplished entirely by plain old neurons. So there seems to be no obvious reason why another sufficiently complex but purely physical system—say, a system made up of copper wires and silicon chips and disk drives and video monitors—couldn't also be conscious. But on the other hand, we feel equally strongly that no amalgam of wires, transistors, silicon, and plastic, no matter what its size or complexity, could move beyond the manipulation of symbols according to their syntactical properties. After all, this capacity for symbol manipulation is exactly what makes computers such powerful and flexible tools. But we know that the semantical properties of symbols can't be discerned from their syntactical properties alone. Thus, this second conviction leads to the conclusion that computers cannot be conscious. So some people find themselves struggling with two apparently contradictory notions, both of which seem obviously true:

1. Because the brain is both conscious and, at bottom, physical, there is no reason to suppose that other purely physical systems, such as computers running certain software, couldn't be conscious.

2. Computers, regardless of their size and complexity, can recognize only syntactical properties of symbols. Knowledge of syntax is insufficient to produce knowledge of semantics. Semantic knowledge is necessary for consciousness. Therefore com-

puters cannot be conscious.

But this apparent contradiction can be quickly resolved. The solution lies in a recasting of the second claim. Physical systems can be thought of at a number of different levels of abstraction: tables, for example, can be considered from the atomic, molecular, carpentry, or furniture levels. Similarly, while computers are traditionally thought of at higher levels of abstraction (i.e., at the level at which programs are run and users interact with the machines), there is no reason for this level to deserve any privileged status over the lower, more mechanical levels of abstraction.[5] So while it is natural to consider a computer to be incapable of consciousness while looking at it at a high level of abstraction, considering the same machine at a lower level—at level of silicon and wires rather than syntax and symbols—reminds us that it is a purely physical system, and as such it may in fact be capable of producing and sustaining consciousness. Recognizing this lower level and the causal powers that lurk there opens at least the possibility of consciousness, and the apparent contradiction vanishes.

Before leaving the topic of the Chinese room, a quick summary is in order. In this portion of the chapter I hope to have demonstrated that the Chinese room argument survives Rey's criticisms about misplaced emphasis on the Turing test, over-reliance on the autonomy of language, and the fallacy of division. I have also tried to clarify a common misconception about the intended conclusion of the argument. Given that most of the complaints made about the argument draw either on the same issues that drive Rey's concerns or on this misunderstanding about the aim of the argument, I believe that the Chinese room is likely to survive any other attacks that have been or will be leveled against

[5]More on levels of abstraction is coming in the second half of this chapter.

it.

Despite all of this, there are those who remain unconvinced by the Chinese room. Some of these people stick to the old claim that consciousness and intentionality can be produced by virtually any type of computer as long as it is running the right sort of program. But about twenty years ago there appeared a new class of machines—computers designed according to so-called "connectionist" principles—which, for various technical reasons, are sometimes thought to be more likely to be able to be programmed for consciousness and intentionality. Let us now look at the details of connectionism more closely, and see what hope it can offer the proponent of computer consciousness.

4.2 Brain architecture: connectionism to the rescue?

Accepting that the mere running of a program is insufficient to produce a fully conscious mind is a good first step toward understanding the relation between the brain and the mind. This realization allows us to bracket off a vast area of research: we can see that all attempts to produce a fully conscious mind through software alone stand no chance of succeeding in any meaningful or interesting way. But the Chinese room is often thought of as addressing only the issue of whether a traditional computer—a von Neumann machine which processes instructions serially rather than in parallel—could sustain consciousness. The argument, or so its critics claim, fails to cover a very different sort of computer architecture known variously as neural net architecture, parallel distributed processing, or connectionism. I believe that because computer architectures of all sorts reduce to Turing machines, and the Chinese room is fully effective against Turing machines, the Chinese room

does indeed weigh in against *any* architecture which ultimately relies on strictly followed algorithmic processes. But rather than showing formally how connectionist machines can be mathematically reduced to Turing machines (this reduction is spelled out in detail in virtually any introductory book on connectionist modeling), I will take a different tack toward deflating the hype of connectionism. Rather than arguing that connectionism cannot sustain consciousness, I will argue something even more basic: connectionism cannot represent mental contents in the same way that the brain does. Now these two issues—the issue of consciousness and the issue of mental representation—are clearly not the same thing, and it is very well possible that a system that represents its mental contents differently than we do might be conscious in just the way we are despite this fundamental difference. But one of the supposed strengths of connectionism is its similarity to the physical structure and layout of the brain, and if I can show that this similarity is in fact not nearly close enough to allow for even a low-level phenomenon such as mental representation, then our faith in connectionism's ability to mimic higher-level cognitive phenomena such as decision-making or consciousness will be severely shaken. This is not the usual approach taken to attacking connectionism, so I will state the goal of this section once more in order to make it as clear as possible. While I think that the Chinese room argument can be adapted to deal adequately with the prospect of connectionist machines producing consciousness, I will try to show that connectionism fails at a much lower, more fundamental level in its effort to mimic the brain. I will show that connectionist machines cannot represent things (such as beliefs, memories, or desires) in the same way that the brain does. My argument, if successful, will not constitute an airtight argument against connectionist consciousness, but it should make

it seem very unlikely.

Looking at the anatomical makeup of the brain, it is not hard to see why connectionist models of cognition are so appealing. Considering that we know the brain to be made up of around 100 billion interconnected but fundamentally unintelligent neurons,[6] connectionism seems an obvious, almost automatic, candidate for describing the representational scheme used by the brain. Many who are unconvinced by the architectural evidence are nonetheless swayed by the ability of connectionist systems to display a capacity for learning, exhibit graceful degradation after sustaining damage, perform successful recognition of distorted or degraded input, and show other qualities characteristic of human cognition. I believe, however, that the question of how closely the brain's architecture is reflected in the architecture of connectionist machines is still very much open. In my analysis of connectionism I will first present a brief explanation of the connectionist model, followed by a discussion of the notion of levels of organization within the brain. I will then illustrate problems faced by connectionist models of cognition that rely on so-called "ultralocal" representation, and present separate problems faced by models that rely on "distributed" representation. Finally, I will provide two additional arguments that are equally effective against both ultralocal and distributed forms of connectionism. I offer these arguments not as concrete proof that connectionism is a false description of the brain's representational scheme, but rather as problems that must be adequately addressed before connectionist models of the brain can be seriously considered.

[6]Current estimates actually range from 85 billion up to one trillion neurons in the average human brain. These estimates are frequently revised (almost always upwards), but the most recent research on the subject that I am aware of is in Pakkenberg and Gundersen [71].

4.2.1 The nature of connectionist models

Since the appearance of James McClelland and David Rumelhart's trailblazing *Parallel Distributed Processing* in 1986, connectionist models of the brain have received a great deal of attention as alternatives to classical models. The classical models of the brain that were used prior to the advent of connectionism describe a brain that processes information serially—i.e., the brain receives a piece of information from the senses or fetches it from memory, processes it, and then moves on to the next piece of information. As mentioned above, this sequential processing is similar to the way that most present-day computers operate. Connectionist models replace serial processing with parallel processing, whereby many pieces of information are acted upon concurrently. Information is represented either by individual nodes or by the strength of connections between nodes, and the processing that occurs makes use of and modifies the weighted connections between nodes. When one node is activated, it sends signals to all other nodes connected to it; whether these signals promote or inhibit activation in these other nodes depends on the weighting of their connections with the original node. A connectionist system's behavior can be modified by changing the number, mapping, or weights of these connections. The name "connectionism" comes from the way in which information is spread throughout the system and the fact that many nodes can be active and acted upon simultaneously.

4.2.2 Levels of organization

Before examining arguments against connectionist models of representation and processing, it is important to make clear the notion of *levels of organization* within the

brain. This fairly common concept appears in many different places in the philosophical literature, but an especially clear presentation is given by Kim Sterelny, who claims that any satisfactory theory of how the brain produces cognition and consciousness must explain these phenomena at three levels: the ecological level, the computational level, and the physical (or implementational) level [102, pp. 42–61].[7] While each level relies somewhat on the level immediately below it (i.e., a theory addressing one level will generally make some minimal assumptions about how the brain operates on the other levels), the levels can be thought of as largely independent. Other philosophers have carved cognition into a more intricate hierarchy of levels, drawing finer distinctions between the functions of each, but Sterelny's three-part division should work fine for our purposes. Theories on the ecological level describe *what* a brain can do without being concerned with *how* it is done. A claim such as "an average person can perform arithmetic, ride a bicycle, and discuss politics," would be cast at the ecological level; the claim describes things a person can do and the types of interaction he enjoys with the outside world. Theories on the computational level present rough algorithms for how skills are learned or used. They dig deeper into the mind and are concerned with sketching out how a brain might go about accomplishing mental tasks. Marr's famous theory of visual field recognition falls largely within the computational level [62], as would theories proposing algorithms the brain might use to perform arithmetic. Finally, physical-level theories are concerned with the finest level of detail within the brain that might be of interest to philosophers of mind or cognitive neuropsychologists. These theories focus on the lowest level of the brain that is still distinctly biological, speculating

[7]The specific distinctions between Sterelny's levels seem to be adapted from Marr's theory of levels as presented in his classic work *Vision* [62, pp. 24–27]. Marr's computational level is equivalent to Sterelny's ecological level, Marr's representational and algorithmic level is Sterelny's computational level, and Marr's hardware level is Sterelny's physical level.

on issues such as the physical layout of neurons but stopping short of making claims about the chemical make-up of the neurons or the laws of physics that govern their constituent atoms. Physical-level theories are perhaps the easiest of the three levels to test empirically, and most people are now quite confident in the standard physical-level description of the brain as a complicated tangle of interconnected neurons. I will not discuss physical-level theories here—because such theories are concerned more with the physical structure of the brain than with the issue of how connectionist models represent things or how well suited they are to sustaining consciousness, they are largely irrelevant to any analysis of cognition (including mental representation).[8] Nor will I be trying to judge whether connectionism can provide answers to ecological-level questions; I will leave this area of study for psychologists. I am concerned here instead with whether connectionism is an accurate computational-level theory. All references in this section to the brain's architecture or capabilities should therefore be understood to refer only to the computational level. Those who argue that we should investigate additional levels that may straddle either the computational and ecological levels or the computational and physical levels may well be right to think that analysis of such levels can prove helpful to understanding cognition and the brain, but they should find that many of the arguments presented here can be adapted to pertain to those levels as well.

[8]Despite the practical irrelevance of physical-level theories to connectionism, the physical structure of computer implementations of connectionist systems are often similar in many respects to the brain's neural structure. However, important structural differences remain. For example, Crick and Asanuma note that actual synapses in the brain show significantly less flexible behavior than most computer-modeled nodes, and that certain types of neurons, such as so-called *veto cells*, are rarely included in connectionist computer models [22].

4.2.3 Connectionism with ultralocal representation

A connectionist model of the brain could store mental objects such as ideas, beliefs, or desires in one of two ways. First, it could assign a single node to each mental representation regardless of the complexity of the representation. The general idea of light bulbs would be assigned to one node, the idea of a particular light bulb would be assigned to another, and the idea that burnt out light bulbs need to be replaced would be assigned to a third. This storage scheme is known as "ultralocal representation." Alternatively, a connectionist brain could use "distributed representation," whereby each mental object is represented by a pattern of activation over a number of nodes (or, sometimes, in the strength of connections between nodes). The general idea of light bulbs may be represented by activating a certain group of five nodes, while the idea of a particular light bulb may be represented by the same five nodes in different states of activation, say, the first four nodes on and the fifth off. In this section I shall examine arguments against connectionist models of cognition that employ ultralocal representation, and in the next section I will present arguments against distributed representation.

It is difficult to see how ultralocal representation could allow a connectionist system to store all the attitudes and ideas that we might expect to find represented in a normal brain. First, I should note that *nodes* in a connectionist model of the brain are meant to correspond loosely to individual *neurons* or groups of neurons in the brain itself. Now, assuming that *every* attitude or fact that the brain represents takes up at least one neuron,[9]

[9]Presumably, many facts that we may claim to know are not actually stored in our minds but are constructed as needed. Certainly we do not store the knowledge that $32,430 \div 6 = 5405$, for if questioned we would have to call upon our stored knowledge of division and compute the quotient on the fly. Even after we disregard facts of this kind, we are left with a staggering number of stored facts to account for.

and assuming further that not all of the brain will be available for storing data—besides the massive number of neurons dedicated to motor skills and sensory processing, many more neurons will be needed to perform support functions related to cognition—it is probably safe to assume that over the course of a lifetime a brain will collect more accessible thoughts than an ultralocally represented connectionist model would be able to store. Even if a connectionist model were given 100 billion nodes—one for every neuron in the brain—that model would very likely, strange as it may sound, fill up. A model of the brain constructed around classical rather than connectionist architecture could avoid this problem by breaking ideas into constituent parts and storing these parts separately. An idea such as *red bicycle* could be stored under classical architecture as a combination of three separate ideas: *red*, the notion of concatenation, and *bicycle*. The idea *red Volkswagen* could reuse the representation *red* and the representation of concatenation, adding only the representation *Volkswagen*. In this way, the use of two additional ideas—the idea *redness* and the idea of concatenation—effectively double the number of ideas that can be represented in a classical architecture brain. A connectionist model that relies on ultralocal representation, in contrast, could only double its number of stored ideas by allotting twice the number of nodes. A model of this type could not achieve the same efficiency by assigning *red* to one node, concatenation to another, *bicycle* to a third, and then activating all three nodes simultaneously, for this would transform the model's representational scheme from ultralocal to distributed. A true ultralocal representational scheme would require separate nodes for *bicycle*, *red bicycle*, *light red bicycle*, *rusty red bicycle*, *my brother's old red bicycle*, etc. Given the extreme inefficiency of the ultralocal representational system, the brain's 100 billion nodes (or neurons) begin

to look more than a little cramped.

The connectionist might try to counter this objection by pointing to the number of facts that we forget over the course of a lifetime; perhaps we simply forget old memories whenever we approach our mental limit. Certainly we forget many things, but given the vast number of complex representations that we do retain, even this does not seem a sufficient response. The bewildering complexity of visual images that we remember vividly, the richness and detail of the countless sounds that are engraved in our aural memory, and the fine nuances of the many aromas that we can recollect would all demand huge numbers of nodes under the ultralocally represented connectionist scheme. The use of symbols and sub-symbols affords classical architecture the vastly increased efficiency that seems to be needed to effectively store all of these representations.

My second argument against connectionist systems that employ ultralocal representation is drawn largely from the work of Jerry Fodor and Zenon Pylyshyn. Although they present several independent arguments against connectionism [34], most of these have drawn quick and convincing responses from the connectionist camp. Only their argument from systematicity still appears to pose a genuine threat. Similar thoughts or utterances seem to share structures in some sense. That is to say, the contents of similar thoughts have an element of symmetry to them. Fodor and Pylyshyn speculate, and they are likely exactly right about this, that the brain exploits this symmetry or "systematicity" to increase its range of thought and expression. If I know how to think and say "I like chocolate ice cream," and I know how to think of and say "coffee ice cream," then I am also able to think and say "I like coffee ice cream." This knowledge would also help me think and say similar

but less closely related things like "I like spaghetti," but probably would not help at all with completely dissimilar thoughts like "My parents live in a red house."[10] Fodor and Pylyshyn argue that the most reasonable way to explain the phenomenon of systematicity is by postulating that mental representations are themselves made up of sub-representations which can be swapped into and out of the larger structure. This component-based representational structure is characteristic of classical architecture but is utterly incompatible with the representations we find in ultralocally distributed connectionist models. Such connectionist representations are coherent, integrated wholes that cannot be dissected into component parts, and therefore cannot be used to explain the obvious systematicity found in human language and thought. This fundamental deficiency in ultralocal representation seems to be a fatal flaw. I know of no response that has been forthcoming from connectionists, and I cannot come up with any response of my own.

4.2.4 Connectionism with distributed representation

The problems of limited storage space and systematicity can be solved by connectionist systems that rely instead on distributed representation, but such systems face problems of their own. Let us first examine how distributed representation can be used to address the problem of storage space. If each node has s states of activation, n nodes could represent only sn different ideas or attitudes using ultralocal representation, but could represent s^n ideas or attitudes using distributed representation. Four nodes with four states of activation each could represent only sixteen ideas under local representation versus 256

[10]However, note that even the last sentence shares a certain systematicity with the previous sentences in virtue of its subject/predicate structure. Except for sentences that serve specialized functions (*"Ouch!"*), the vast majority of sentences exhibit the same symmetry of construction. Systematicity in language (and probably in thought) seems extraordinarily deeply ingrained.

ideas under distributed representation. This advantage of distributed representation grows exponentially as the number of nodes dedicated to representations increases, and also grows rapidly as the number of activation states for each node increases. This scheme seems more than sufficient for human storage needs; if only one tenth of the brain's neurons are dedicated to storing ideas and attitudes, and if each neuron is capable of only two states of activation (and neurobiological evidence indicates that at least this latter figure is extremely conservative), the brain could accommodate approximately $2^{1,000,000,000}$ representations. That should hold us for a while.

The problem of systematicity also appears to be solved by distributed representation. Sterelny makes exactly this point, claiming that systematicity is possible in any representational scheme that supports what he calls "micro-features." [102, p. 186]. Spreading a mental representation over a number of nodes is one way to provide such support. To use a very crude example, a twenty-node representation of my kitchen could presumably have twenty micro-features (or significantly more if each node has several possible states of activation). One of the twenty nodes might indicate the presence of a sink, another a refrigerator, and another a dishwasher. This would seem to allow for systematicity: separate representations of kitchens that differ only with respect to a few micro-features would share very similar representations. Along the same vein, a twenty-node distributed representation of the attitude *I like chocolate ice cream* could be slightly modified—say, turn off node 17 (corresponding to the micro-feature *chocolate ice cream*) and turn on node 18 (the micro-feature *coffee ice cream*)—to form the similar idea *I like coffee ice cream.*

This use of distributed representation to solve the problem of systematicity seems

plausible, but it raises a new problem: it does not allow for plasticity of micro-features. A distributed representation of a kitchen would have room only for a finite (and preëstablished) number of micro-features, and it is difficult to see how one could continue to add micro-features as new discoveries about the kitchen are made. If additional nodes are added to the representation to make room for new micro-features, this change of the representation's size would seem to require the re-weighting of all connections leading to or from the representation. That is, adding a new micro-feature would effectively result in an entirely new representation, and the delicately balanced system would, it seems, have to be completely re-trained on this new representation. Realizing that arguments from introspection are often suspect, I would argue that this mental maneuver seems wildly unrealistic—adding new features to a representation in no way seems to involve reconstruction of the entire representation. In contrast, the symbols and sub-symbols of the classical model would not only allow the mind to represent any arbitrary number of micro-features, but could accommodate micro-features *of* other micro-features. Not only could a classical representation of a kitchen indicate the presence of a refrigerator, but it could also indicate which corner of the room it is in, what color it is, and how loud its compressor is. The recursive capacity of classical architecture allows it to generate representations whose depth and complexity are limited only by the finite size of the brain. Unless the connectionist can propose an equally rich form of representation, classical architecture would seem to enjoy a tremendous advantage.

Let us return to the issue of limited storage space for a moment. While we have seen that the use of distributed representation does allow for more storage space than

ultralocal representation, it raises another, somewhat related problem: if different mental objects reside within the same set of neurons (where different patterns among those neurons represent different ideas or attitudes), it is very hard to see how our minds could think of two such objects at once. Calling to mind a specific representation would set the neural network in a specific pattern. Thinking of certain other representations (namely, those that share nodes with the first representation) would require a new pattern to form among those same nodes. Destroying the first pattern should then cause the first idea to drop from consciousness. Because connectionist models require separate representations even for very closely related ideas or attitudes, one should not be able to think simultaneously of, say, one's fear of going to the dentist and one's fear of heights. Yet it seems that we frequently do exactly this; how else could we compare the two fears when, say, trying to decide which is more debilitating? How could we compare any two things at all?

The connectionist might respond with an argument drawn from psychological evidence. That most people can store roughly seven items in short term memory suggests that the brain may be divided into seven subsets of neurons, each capable of functioning independently of the others. Each subset would still be large enough to store a huge ($2^{1,000,000,000 \div 7}$, if there are two states of activation per neuron) number of ideas, and up to seven different ideas could be considered concurrently. But assuming that each idea is stored in only one subset of the brain, the connectionist must then explain how we can use multiple instances of an idea. Thinking *the dog bit the cat* would pose no problem to a brain that stores *dog* in only one location, but such a brain would have difficulty forming the thought *the dog bit the dog.* Furthermore, there would still be conflicts between ideas

represented within the same subset of neurons; if our ideas are spread evenly among the seven subsets, we would expect any given idea to conflict with one seventh of all of our other ideas. The connectionist could get around these problems by suggesting that there are copies of each idea, with one copy in each neural subset, but this seems an inadequate explanation given how easily we forget ideas. On this model, the process of forgetting would presumably have to be fully synchronized throughout all seven neural subsets, so that forgetting an idea in one subset would force all parallel ideas in other subsets also to be lost. It seems unlikely that this would happen with perfect consistency. Similarly, it is hard to imagine the neural connections to and from all copies of an idea being evenly and accurately maintained over time. As new associations with an idea are formed—say, I learn that a tomato is a fruit rather than a vegetable, so I cut the tomato-to-vegetable connection and form a tomato-to-fruit connection—it seems unlikely that all copies of the idea would be updated perfectly. We would expect occasional discrepancies to form between parallel ideas, which would then produce strange and unpredictable effects. Some copies of my idea *tomato* might inadvertently be left with connections to my idea *vegetable*, and if I were asked whether tomatoes were fruits or vegetables, my answer might well depend on which copy of *tomato* I access. So this use of neural subsets to solve the problem of conflicting ideas seems less than satisfactory.

I can think of two other responses to the problem of conflicting ideas that the connectionist could propose. First, he might admit that we do sometimes experience this effect, perhaps offering the Necker cube as an example (see Figure 4.1). We are generally able to view only one orientation of the cube at a time; very few people can see both

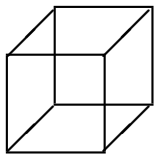

Figure 4.1: Necker cube

simultaneously. This could indeed be caused by conflicting connectionist representations, but it is equally possible that the phenomenon stems from a natural limitation of our biological mechanism of visual processing—the problem could arise long before connectionist elements of the brain even come into play. The connectionist might also suggest that our apparent ability to consider two ideas at once is merely illusory, and that we are in fact switching quickly between the two. This response may solve the problem, but only at the expense of abandoning the essential nature of connectionism. To implement this solution, the connectionist model would have to shed its connectionist identity and take on a suspiciously classical form. Switching quickly between two ideas sounds like a task much better suited to a mind based on serial classical architecture than on parallel connectionist architecture. Also, one might question how a connectionist mind would know to return to the first idea once the first idea had been supplanted by the second. As soon as the mind dedicates itself completely to the second idea, it would seem to lose track of its past train of thought; by focusing on the second idea it would forget that it needed to return to the first. If the mind were able to maintain a memory trace of some sort—say, a mental bookmark or pointer to the first idea—then we might ask why it could not simply maintain the first idea itself. So this response seems to me to be a non-starter.[11]

[11]In the interest of completeness, I should note that there is another, fully independent objection to

4.2.5 Additional objections

My two final objections to connectionism apply to both ultralocal and distributed versions. The first objection focuses on the rate at which connectionist systems learn. Computer models built on connectionist principles have had great success in learning tasks through back-propagation and outside tinkering of connection weights, and this capacity to learn is one the most persuasive arguments in favor of a connectionist as cognitive architecture. However, the strength of this evidence is tempered by the extremely slow rate at which these models learn. With only a few thousand nodes, competent character recognition or image analysis is achieved only after many thousands of training cycles. In general, the more complex the task, the more nodes are needed. The larger the number of nodes involved, the less effect each training cycle has on the weights of each node's connections, and the more training cycles are required. One can scarcely imagine how many training cycles it would take to fully train a 100 billion-node connectionist machine to accomplish all the extraordinarily complex tasks that human brains routinely perform with their 100 billion neurons. The connectionist models that have been successfully constructed accomplish *very* specific tasks, and even these narrowly focused models can require tens of thousands of cycles to train. If we were to mimic human intelligence by constructing separate connectionist systems that each accomplish a single, distinct task routinely performed by humans, we would still need a way to combine this enormous collection of independent systems into a unified whole. Simply stringing together a series of sub-systems to form a complete model

distributed representation that I will not discuss in detail here. This is the objection that certain logical relations that we find quite natural to think about are difficult for distributed-representation connectionist systems to derive or use. For example, if we know that a table has a leg, and that that table is in a room, then it is obvious to us that the leg is also in the room. Many connectionist systems have great difficulty arriving at this same conclusion.

of the brain might well require more than the 100 billion nodes we have to work with. Blending the sub-systems together in order to share nodes would probably be possible but would likely require an unrealistic number of cycles to train.

Although learning obviously continues to some extent throughout one's life, humans already grasp many essential skills after only ten years of life and are considered to reach maturity after twenty. Assume that one third of our lives is spent in sleep. The mind continues to work during sleep, but I maintain that very little learning takes place at such times; other than the pressure of the mattress and perhaps the ticking of an alarm clock, there is very little input from the outside world to train the mind. This means that after only 60,000 hours of wakeful training, the 100 billion neurons of a ten year-old child's connectionist mind would have mastered such diverse skills as the production and comprehension of spoken and written language[12] (quite possibly in three or four languages); the learning and practice of thousands of quite complicated motor skills; the learning of rudimentary rules of social interaction; the comprehension of simple mathematical concepts; the ability to recognize thousands of different foods by their taste; the ability to mentally map a geographical area; the ability to dissect, process, and recognize complex visual fields such as human faces; and the ability to recognize, appreciate, and produce music. Considering that actual neurons in the brain are known to fire, on average, one thousand times less frequently than simulated neurons in connectionist models implemented on supercomputers (100 firings per second for the brain versus a rather conservative estimate of 100,000 firings per second for the computer), this feat of human learning becomes all the more remarkable

[12]Certainly great debate exists over how much language learning actually occurs during childhood, but even in the best-case scenario (say, if the brain were born with some degree of linguistic knowledge preweighted among its connectionist nodes, as Chomsky might suggest were he to imagine a connectionist brain), the brain would nevertheless have to learn a great deal of vocabulary, spelling, and the like.

[22, p. 366]. Again, this disparity in rates of learning makes connectionism look distinctly unqualified to be a candidate for the computational-level architecture of human cognition.[13]

A final objection to connectionist models is best stated as a question: if the computational level of my brain is connectionist in nature, why don't random ideas and attitudes present themselves as my brain settles into stable states? Each state of the neural network would have to correspond to an idea or attitude, or to some support function necessary for the brain's operation but not noticeable to the conscious mind, or to nothing at all. As data enter the brain via the senses, the brain would have to shift toward a stable state, remain briefly in that state, and then move toward another stable state as more input arrives. We must assume that stable states could be quite different from each other and that the overall pattern of the network could therefore change significantly in moving from one state to the next. In this process of shifting from one stable state to another stable but distant state, it seems almost certain that the neural network would pass through many intermediate states that represent unrelated ideas or attitudes. Why don't these strange and random notions occur to us whenever we think? When my mind shifts from thinking of pizza to thinking about the stack of exams that I have postponed grading, what prevents thoughts about my brother, snippets of my favorite symphony, or scenes from last night's movie from flashing into and then out of my consciousness as the states that represent them are briefly visited? Random thoughts certainly do occur, but they generally arise in my brain when

[13]Figures presented here are based largely on a connectionist model developed by Rumelhart and McClelland that shows human-like behavior in learning the past tenses of verbs [87]. Steven Pinker and Alan Prince argue that the success of this model, considered one of the most thoroughly implemented and successful connectionist models, is due largely to the learning set on which the network was trained. Pinker and Prince claim that *even more* training cycles would be needed for connectionist models such as Rumelhart and McClelland's to perform well over more general input (in this case, a wider range of verbs) [78]. This suggests that the rate-of-learning problem may be even more daunting than I make it out to be.

I am relaxed and thinking of nothing in particular. When thinking hard about a specific problem, I tend to have thoughts related only to that problem. How, then, can my mind stay so focused?

Four possible connectionist responses to this worry suggest themselves. First, different thoughts about a single problem might have similar representations, so shifting from one such thought to another would require only a short journey through the landscape of representations and neural configurations. Second, the occasional appearance of random thoughts might help to explain creativity. Third, the brain might throw up a filter to trap stray thoughts and block them before they reach consciousness. Finally, perhaps only stable states last long enough to thread their way into our consciousness.

The first solution seems implausible considering the wide variety of sensory input that constantly flows into our brains. Of the countless sights, sounds, and smells that bombard our senses while we focus on thoughts unrelated to the sensory input, surely some of these would jolt our brains into registering unexpected ideas. Unless we think only and completely of the data we gather from our senses—say, think only of a blank white wall while looking at a blank white wall—the act of sensing anything at all ought to jog our minds enough to cause random thoughts. Furthermore, it is difficult to imagine being able to think constantly and exclusively of what we sense. This task seems impossible even when limited to vision. Before turning my head I would have to think about what I expect to see from the new perspective and then think of the new view at precisely the moment I shift my gaze. I would also have to follow the same practice for every sound, smell, taste, and feeling that is conveyed through my senses. I could never allow anything to surprise me.

And even if we were able to exercise such strict control over our thoughts, it is obvious that we rarely, if ever, actually do this.

The notion that random thoughts do occur and that they form the basis for creativity deserves more careful consideration, but unless classical architecture is unable to explain creativity, this need not be a deciding factor in favor of connectionism. The proposal of a mental filter also seems weak. We do appear to use similar filters to block out sensory input—it is easy to become so engrossed in one's work that one fails to notice the ticking of a clock or the dimming of the room due to the setting sun—but it is much easier to imagine a filter preventing information from bridging the gap between the senses and the brain than it is to imagine a filter preventing thoughts already within the brain from reaching consciousness. If the ideas or beliefs of which we are conscious are simply those that are represented by the arrangement of active neurons, it is hard to see how these ideas or beliefs could be blocked at all. Would blocking not result in great gaps in our thought, or periods in which *no* representation presents itself to our consciousness? Empirically, this seems not to be the case. One finds it hard to imagine what it would be like to think of absolutely nothing at all. Furthermore, it is not clear that a connectionist model would even be able to implement such a filter. That would require some sort of overseer operating independently of the connectionist system, for something outside the system would have to be responsible for activating the filter when needed and shutting it off when all random thoughts have safely passed. An overseer of this type is much more likely to exist in classical models of the brain, for classical models require just such an entity to manipulate symbols and keep the "program" of the mind running smoothly.

The fourth response may be the connectionist's best bet: perhaps only stable states last long enough to register on our consciousness. I cannot offer a definite argument against this suggestion, but I can surmise where such an argument might be found. We know that human sensory perception is composed of discrete input signals that are somehow transformed into continuous sensations: the eye registers 30 to 50 frames per second, and nerves from the fingertips, taste buds, ear, and nose fire a finite number of times each second. No doubt experiments have already been performed to determine the shortest amount of time a stimulus must be present in order to register on any of the five senses. Assuming that the brain's neurons fire at most 100 times per second, and assuming that no echo of the stimulus lingers in later states, if experiments were to show that stimuli that last less than $\frac{1}{100}$ second *can* be perceived, and if at least some cases were found in which further stimuli that immediately followed the original stimulus failed to mask the original stimulus, then we might be led to believe single "cycles" of the brain's neural states may register on our consciousness.[14] This in turn would suggest that perception of the temporary states the brain briefly passes through on the way from one stable state to another also ought to reach our consciousness. This line of argument is of course highly speculative, but I can see no reason why it is not empirically testable. However, I leave that task to experimental psychologists and neurobiologists.

[14]This argument also assumes that the brain shifts from one state to the next in regular cycles. That is, all of the neurons fire at the same time, rest, and then fire again $\frac{1}{100}$ second later, resulting in discrete transitions between states. If neurons were to fire asynchronously then states could last for much less than $\frac{1}{100}$ second— if no two of the brain's 100 billion neurons fire simultaneously, states would last only one 100-billionth of a second! In fact, the existence of *grand mal* epileptic seizures brought on by the synchronous firing of massive numbers of neurons might be taken as evidence against the very assumption I am making.

4.2.6 Review and summary

I've made a number of independent attacks on connectionism in this section, so let me review them briefly. Connectionist models of cognition that rely on ultralocal representation seem hampered by a lack of storage space, and their inability to accommodate the obvious systematicity found in human thought is troublesome. Models that use distributed representation solve these problems, but they allow insufficient flexibility of micro-features within representations and they provide limited capacity for multiple concurrent thoughts. Even if the problems faced by ultralocal or distributed representation are overcome, the proponent of connectionism must still account for the apparent disparity between human and connectionist rates of learning, and must provide a plausible explanation for how a connectionist model could block spurious mental states from reaching consciousness. While it is hard to deny that the massively interconnected neurons that make up the physical architecture of the brain suggest that higher levels of the brain's organization might be connectionist in nature, these arguments show that it would be at worst misguided and at best premature to endorse such a conclusion.

4.3 Where do we go from here?

Machine consciousness appears to be in dire straits. We first looked at some of the difficulties that attend the claim that software running on a classical serial-processing von Neumann machine could sustain consciousness, and then we saw that there are equally difficult problems for models based on connectionism. These two theories pretty much exhaust the major architectural options that are taken seriously by philosophers, cognitive

scientists, and anyone else who has been trying to endow machines with consciousness and human-like intelligence.

There are in fact other, less popular and less well received approaches that are being investigated today, but they can be dismissed rather quickly. Many simply reduce to one of the two schemes we've discussed. For example, some of the earliest, most thoroughly worked out, and certainly most famous attempts at producing artificial intelligence—Herbert Simon and Allan Newell's series of increasingly mature "physical symbol system" models—are just elaborate versions of the classical model.[15] And while functionalist solutions to the mind/body problem are not always applied directly to the issue of machine consciousness, most of them could be quite easily pressed into service to lend support to the classical model as well. Connectionist models are generally labeled as such (connectionism drawing a certain amount of cachet these days from its reputation of being cutting-edge and higher than high-tech), but some early, pre-connectionist models such as Marvin Minsky's "society of mind" [67] could perhaps also be seen as fitting into the connectionist mold. In fact, connectionism receives some form of support from virtually any theory that proposes a collection of separate but interconnected components that are fundamentally dumb but which combine to produce spectacular higher-level phenomena such as consciousness or cognition. Daniel Dennett's "multiple drafts" theory of consciousness is probably both the clearest and the most detailed example of this [25], but this general strategy is actually a rather common approach.

While I don't mean to suggest that these are the only two avenues open for the

[15]Newell and Simon practically invented the classical model of cognition with their early work on artificial intelligence. For a clear and simple exposition of the theoretical underpinnings of their model, see their Turing lecture from 1975 [70].

modeling of cognition and consciousness, they are certainly the best known and, frankly, the only options taken at all seriously by most researchers. Alternative models tend to be short on carefully worked-out detail and long on speculation. They are often the products of long-term projects kept alive by small groups of committed researchers who have invested too much time and energy into their theories to be able to write them off. For example, Rodney Brooks supports an interesting cognitive model which seems to do away with representation entirely [11]. He somewhat cryptically claims that his system uses the world as a model of itself, and so needs no mental objects of any kind. The lack of explicit, localized representation gives this theory a distinctly connectionist flavor, but Brooks is adamant that it has neither sufficient numbers of interconnections nor sufficient homogeneity in the structure and behavior of its nodes for it to be considered a connectionist model.

Or consider Timothy van Gelder's "dynamical model" for producing a mind [110]. He dispenses with representations just as Brooks does, but he also eliminates explicit rule following by envisioning the brain (on some unspecified level) as a series of coupled dynamical systems. This means that each system influences the behavior of the others through self-regulating mechanical (rather than computational) means, much like the value of each variable in a differential equation directly affects the value of certain other variables in that equation.[16] This model is *sui generis* in the world of cognitive science, lying well outside of both classical and connectionist camps. The latter fact is not immediately obvious,

[16]This concept is difficult to convey succinctly, but Van Gelder gives a nice example of a working dynamical system when he explains how James Watt's eighteenth century steam engine has its speed governed by an ingenious dynamical mechanism. As the engine spins faster, it swings a weighted arm further outward, which in turn opens a valve that releases steam pressure which then slows the engine speed. This results in a constant engine speed even as the amount of steam flowing through fluctuates. Van Gelder's claim is that this elegant system is completely computation- and representation-free. There is of course room to argue with him on this point, given that representations and computation seem to be *imposed* on a system by outside observers rather than *intrinsic* to that system. See Searle [95, ch. 9] for more on this.

given that so-called "back-propagating" or "recurrent" connectionist models—i.e., models that have connections running backwards from downstream nodes to upstream nodes through which the data have already passed—are strongly dynamical, since each node has the potential to affect all other nodes (or at least many other nodes, depending on the exact configuration). But this is not the only way to set up a connectionist model: "feed-forward" configurations, in which data passes only in one direction, are often preferred on the grounds that the brain itself uses very little (if any) back-propagation. In any case, Van Gelder is careful to distinguish his model from connectionism by saying that "nodes" in his model, such as they are, are not sufficiently homogeneous in structure and there are not enough interconnections between the nodes for the model to count as an example of full-fledged connectionism. In sum, I consider this an intriguing theory given the success of dynamical systems in solving real-world engineering problems, but it is not at all clear how exactly the general concept could be applied to the brain, or how it would then help us understand consciousness or the full range of cognitive tasks. Nevertheless, that dynamical systems have been used somewhat successfully to mimic small components of cognition (such as James Townsend's "Motivational Oscillatory Theory," which models human decision-making processes [107]) suggests that it is worth keeping an eye on developments in this area.

There are yet other options available. Although these approaches seem to have fallen out of vogue recently, there was a period not so long ago when many attempts to produce consciousness relied on the closely related concepts of *embodied artificial intelligence* (EAI) or *artificial life* (AL). EAI emphasizes the importance of embedding a computer mind—whether classical, connectionist, or other—into a robotic body through which it can perceive

objects and initiate actions. The theory is that without such a body, the artificial mind's inability to perceive or act means that real semantic content never has the opportunity to enter the system. Not surprisingly, EAI served as the impetus for the Robot Reply to the Chinese room argument, as discussed earlier in this chapter. AL is a strategy in which only the most basic skills and knowledge are programmed into a computer brain, after which the model builds additional, higher-level skills naturally and autonomously via its interaction with the outside world. The long-running "Cog" robot project at M.I.T. is perhaps the best-known example of this [1, 12].

The list of proposals for brain models, and hence for producing consciousness in machines, goes on. Other candidates include Bernard Baars' global workspace [6], Francis Crick and Christoph Koch's theory that consciousness arises from cortical neurons firing synchronously at 40Hz [21, 23], Roger Penrose and Stuart Hameroff's claim that the origins of consciousness are found in the quantum-level behavior of microtubules within each human cell [76, 77], Gerald Edelman's concept of reëntry mapping [32], and David Chalmers' dualistic functionalism [15]. All of these theories are interesting and unique, for none of them falls cleanly into either classical or connectionist camps. Of these, I feel the most sympathy toward Chalmers' theory simply in that it reëmphasizes both the central role that qualia play in the production of consciousness and the irreducibility of qualia, but those features of his theory are probably best carved off and used independently, as his arguments for functionalism and (especially) panpsychism are not entirely satisfying.

None of these views has drawn very wide acceptance, either because no one quite understands the models involved, or else because no one outside of the immediate research

groups sees any reason to think that they have any legitimate chances of success. In any case—and this is one of the main points I want to make in this chapter—I contend that none of these attempts to either model consciousness in the abstract or to produce consciousness and human-like cognition in machines has yet yielded any truly significant results. But this claim raises two interesting and important questions. First, what sorts of results would we be looking for, anyway? That is, how would we ever be convinced that consciousness exists in some non-human system? And secondly, what would be the implications of finding such consciousness? I will focus on these and other related questions in the next chapter.

Chapter 5

Implications of consciousness

Now that we have a working definition of consciousness and a sense of how difficult it would be to produce artificially (i.e., through non-biological means), we might wonder what its value is. Why do we care about it, and how should its presence or absence in a particular system affect our behavior toward that system? These are enormous questions, but I want to try to make some limited progress toward answering them here. However, before we talk about the significance of consciousness, I should address two issues that are in some sense prior. First, what does my notion of consciousness tell us about the mind/body problem? And second, what strategy should we use in looking for it in non-human systems?

5.1 Consciousness and the mind/body problem

One of the most important implications of my definition of consciousness is that the very presence of consciousness demonstrates that the most popular solutions to the mind/body problem are all false. I should emphasize that this is certainly not an original

claim on my part—Searle [95, chs. 1, 2] gives a particularly clear and trenchant presentation of similar material—but given the surprisingly small number of philosophers who accept the claim, I think it is worth repeating. However, because the point has been discussed eloquently by others, I will keep my comments on the subject extremely brief.

The mind/body problem is one of the very oldest problems in philosophy. Most of the solutions that have been proposed over the centuries fit into six general categories: dualism, idealism, behaviorism, identity theory, functionalism, and eliminativism. Dualism takes consciousness very seriously, but the theory faces a host of well-known problems that I will not dwell on here.[1] Idealism—the position that the world consists entirely of ideas within one or more minds—was of course a central element of the philosophies of Berkeley, Kant, and Hegel (among others). Although probably the dominant philosophical worldview at the beginning of the twentieth century, idealism finds very little support among contemporary analytic philosophers. The remaining four theories are currently jockeying for position, with functionalism probably in the lead at the moment. However, all four theories are wrong for the same simple reason: they all fail to account for qualia-based consciousness. I will give a very short overview of each of these four theories and then explain why I claim that they all neglect this most fundamental aspect of human experience.

[1] OK, I'll dwell a little. Dualism's main challenges include solving the problem of mind/body interaction (how does one influence the other?), explaining how souls are created and bound to bodies, and dealing with the fact that Occam's razor would like to pare the world's ontological taxonomy down to just one (presumably physical) type of substance. Very little, if any, progress has been made on these problems since dualism was presented in its modern form by Descartes [27]. However, the theory is not without its supporters: see John Eccles [79, part II] for a Christian-influenced defense of dualism, or David Chalmers [15] for a more recent and decidedly agnostic argument in its favor.

Behaviorism

Behaviorists claim that mental states just are behavior.[2] For example, the mental state of preferring Saabs over other cars consists in visiting a Saab dealer first when car shopping, washing one's Saab more often and more carefully than one washes one's Toyota, or saying the words "I prefer Saabs over other makes" (or at least being *disposed* to act in these ways if prompted appropriately). The only way that a mental state manifests itself is through the behavior of the person who has that mental state—there is nothing else going on inside the subject's head.

This view of the mind has its origin in the logical positivists' quest to make all claims about the world testable. By pulling mental states out of the head and into the external, physical world, claims such as "I am in pain" or "I want to go to Florida" do indeed become testable, and psychology gains a measure of respect as a scientific project. But this comes at too great a price, for behaviorism denies the aspect of mental states that seems most "mental": the *feel* of being in a particular mental state. Behaviorism does away with the "what it's like" component of a mental state, thereby sabotaging any chance it might have at providing a satisfactory account of the mind. In short, it solves the problem of mind/body interaction by getting rid of the mind, and in doing so it bluntly denies the existence of qualia-based consciousness. Perhaps this is why very few philosophers today explicitly embrace this theory (Dennett probably being the most conspicuous exception, and even he seems reluctant to call himself a behaviorist).

[2]See Gilbert Ryle [88] for a well-known explanation and defense of the theory. His 1949 book *The Concept of Mind* is a paragon of clarity and his arguments are very strong—*almost* strong enough to make the position sound plausible. For a more succinct exposition of the view, try Carl Hempel [46].

Identity theory

Identity theories emerged in the 1950s as a response to less-than-satisfying behaviorist accounts of the mind.[3] They come in two distinct flavors. Type-type identity theory claims that there is a particular brain state that is identical to a given mental state, and this brain state remains constant from person to person. So if my feeling angry at the I.R.S. is identical to neuron number seventeen in my brain firing at 24Hz, then your feeling angry at the I.R.S. is identical to neuron number seventeen in *your* brain firing at 24Hz. Token-token identity theory retains the idea that mental states are identical to physical states, but allows these physical states to vary from person to person. This has the advantage over type-type identity of allowing people with different neurological configurations (or species with entirely different kinds of brains) to have the same mental states, which is a possibility that we surely want to accommodate. There are a number of strong objections—both technical and common-sensical—to these theories, but the only one relevant to this discussion is essentially the same objection we used against behaviorism: in equating a mental state with a physical brain state, these theories leave no room for qualia. Once you've said that hunger is just the firing of neuron number 413, there doesn't seem to be any place for qualia. Note that identity theorists are not claiming that hunger is a real phenomenal feel that happens to occur whenever neuron number 413 fires. If that were their story, then identity theory would be perfectly compatible with qualia-based consciousness. Rather, they are saying that *there is nothing to hunger* other than the firing of neuron number 413, which is a position that clearly excludes any notion of qualia. So identity theories do not just fail to

[3]See David Armstrong [4] and J. J. C. Smart [101].

explain consciousness. Rather, they clearly deny the very possibility of the phenomenon; they say that mental states are really physical states, and that is all they are. By my lights, this claim alone is enough to disqualify identity theories. Once again, a solution to the mind/body problem proves unsatisfactory because of a failure to allow for consciousness.

Functionalism

I have already talked a bit about functionalism in chapter four's discussion of machine consciousness, but it is worth reviewing the basics of the theory here. It emerged in the 1960s, although it is unclear whether Hilary Putnam or David Armstrong deserves the credit (or blame) for developing it.[4] Functionalists claim that mental states are identical to physical states, but they differ from identity theorists in that they identify mental states by looking at the *relations* those mental states have to other mental states of the system, to the system's sensory input, and to the system's behavior. A physical state counts as the mental state of experiencing pain or thinking about the Pope in virtue of the function it plays in the overall mental life of the organism. The most significant advantage of functionalism is that it allows for so-called "multiple realizability": the same functional organization can occur within many different physical systems. Thus functionalism avoids the charge of neural chauvinism to which identity theories are susceptible, by allowing for the possibility of a human, a robot, a Martian, or any other appropriately organized system to feel pain or to think of the Pope.

However, functionalism faces the same problem that doomed behaviorism and

[4]The earliest reference to functionalism that I can find is in endnote two of Putnam's 1965 article "Brains and Behavior" [80], where he cites an earlier, apparently unpublished article that may or may not have been written by him. Ironically, Putnam has since abandoned the theory he was responsible for popularizing and is now one of its most prominent critics.

identity theory: it fails to make room for qualia. This weakness is revealed by the well-known inverted spectrum problem, one of many arguments that have been made against functionalism.[5] The mental state produced in you when you see a green traffic light may play exactly the same role in *your* mental life as the mental state that is produced in me when I see the same traffic light plays in *my* mental life. Because these two states play the same functional role in each of us, the functionalist sees them as type-identical mental states. But it seems quite possible that the light could *appear to be different colors* to the two of us, which suggests that the functionalist has left something out of his account of mental states. What he has omitted is, of course, the qualia that accompany all mental states. Once again, an inability to accommodate qualia casts serious doubt on a proposed solution to the mind/body problem.

Eliminativism

Eliminativism is a slightly different kind of theory than those we have already discussed, in that it is perhaps more clearly situated in the realm of psychology than philosophy; it is more a theory of how we ought to think about our mental states and what terms we ought to use to describe them than it is a theory of how those states relate to the brain or body. Nevertheless, it is worth looking at here because it is commits the same error that I have accused the other theories of making.

The best known version of eliminativism is probably eliminative materialism, a theory that Paul and Patricia Churchland have dedicated the better part of their careers to developing and defending.[6] Eliminative materialism does away with our standard way of

[5] See Block [7] for a catalogue of such objections.

[6] Paul Churchland's *Matter and Consciousness* [19] provides a quick introduction to the theory, while his

talking about mental states by arguing that terms like "belief" or "desire" are elements of a "folk psychology" that has long since outlived its usefulness. They argue that these terms are derived from an archaic, stagnant, and unilluminating project to understand the mind, and that we should dispense with them in favor of a detailed neurobiological understanding of the brain. So instead of saying that Jones hopes for rain and thinks that Chicago is in Illinois, we should simply say that neuron 413 in his right parietal lobe and neuron 12 in his left frontal lobe are firing at 35Hz. By eliminating talk of mental states entirely, we will be able to focus on understanding how the brain works rather than on how a set of poorly defined and impossible to observe mental states interact and produce behavior.

It is important to understand that eliminative materialism does not explicitly deny the presence of qualia. Rather, it argues that we should shift the focus of our research away from qualia and toward the brain itself. We can learn about the mind only by identifying and understanding the neurobiological underpinnings of our mental states rather than by paying attention to the qualitative features of those states; studying qualia will simply not prove helpful or interesting.[7] While it is certainly true that we have much to learn about the brain, and any information we can glean about its inner workings will indeed contribute to our understanding of the mind, most contemporary philosophers seem to agree that eliminative materialism pushes this line of thought too far. The Churchlands (and others) do present intriguing arguments in favor of this theory, but the fact that it downplays one of the most common and powerful aspects of human life—the experiencing of qualia—makes it difficult to accept fully. Mental states such as beliefs and desires have

The Engine of Reason, The Seat of the Soul [20] gives a more detailed defense of the view

[7]Perhaps I am actually being too charitable toward eliminative materialism. Brian McLaughlin, for one, seems to understand it as doing away with qualia and consciousness entirely: "Patricia Churchland thinks the appropriate response to consciousness is eliminativist: there is no such thing." [66, p. 605]

distinctively qualitative components, and dispensing with "high-level" talk that involves terms like "belief" and "desire" in favor of "low-level" talk of neural configurations and firing patterns would seem remove qualia entirely from our analysis and understanding of the mind. This feature of eliminative materialism makes it very difficult to accept.

In this section I have tried to show in the most cursory of fashions that qualia-based consciousness is not compatible with some of the major solutions to the mind/body problem. I am not, unfortunately, able to propose a constructive solution to the problem in this dissertation. There are a number of options currently on the table that take consciousness seriously (John Searle's "biological naturalism" [95], Colin McGinn's depressingly defeatist "mysterianism" [65], and David Chalmers' functionalist/dualist hybrid [15] spring to mind), but no theory has as of yet attained anything approaching general acceptance.

5.2 Recognizing consciousness

> *[T]he mind consists of qualia, so to speak, right down to the ground.*
>
> John Searle [95, p. 20]

As I have argued throughout this dissertation, Searle is exactly right on this point. And if the mind and consciousness really are qualia all the way down, that tells us something about how to detect minds and consciousness in systems other than ourselves. That qualia are intrinsically private eliminates the possibility of direct access to other systems' conscious states—a fact that explains why the problem of other minds remains frustratingly intact after thousands of years of efforts to unravel it. We are left, then, only with behavioral and structural evidence to test for the presence of consciousness in other systems. Behavioral evidence is clearly not enough. To return to a point made in my review of Rey's arguments

against the Chinese room, the Turing test has long fallen out of favor among cognitive scientists and philosophers of mind simply because we now realize that behavior is an insufficient indicator of the presence of consciousness or even intelligence. We were hugely impressed by the cleverness of IBM's engineers as we watched their chess-playing computer Deep Blue beat world champion Garry Kasparov in May of 1997, but very few of us considered Deep Blue to be conscious; it just seems extremely unlikely that there is anything at all that it's like to be a massively parallel IBM RS/6000 supercomputer. Or to restate the point slightly differently: no matter how human-like the display of logic or reasoning, no matter how complex the mechanical process that underlies the sophisticated behavior we observe in any system, that behavior does not by itself provide a sufficient reason to believe that that process is accompanied by the experiencing of qualia. Human-like behavior is simply not a sufficient condition for consciousness as I have defined the term, and this fact is obvious to virtually anyone who gives the issue more than a moment's consideration.[8]

But am I rash to make such a strong claim—am I dismissing behavior too quickly? After all, the traditional way to discuss the function of consciousness is to propose a long list of actions that would seem impossible for an unconscious being to perform. Well, if we look carefully at some of the most common elements of such lists we'll see that in each case the tasks can be performed just as well by thoroughly unconscious machines. My responses to

[8]While conspicuous exceptions include those computer scientists or science-fiction fans who *want* computers to be conscious, even popular literature sometimes gets this point right. Consider the protestations of EDGAR, a thinking and behaving computer program in Astro Teller's *Exegesis*, a recent addition to the string of hyperintelligent-computer-threatens-mankind novels that date back at least as far as Arthur C. Clarke's *2001: A Space Odyssey*. The author is clearly getting at the distinction between human-like verbal behavior and the experiencing of qualia: "There is so much I now believe I will never comprehend. I understand 'freedom' and 'intention' and 'quickly.' How can I appreciate 'blue' or 'heavy' or 'pain' or 'musical'? I can not see, lift, feel, or hear" [106]. Of course, this passage suggests that EDGAR is denied qualia by virtue of being disembodied, but I hope that my arguments in chapter four have shown that connecting a program to robotic limbs and artificial sense organs hardly guarantees that that program will experience qualia.

these claims will be very short, but they should be sufficient to demonstrate the significant problems that they face.

Consciousness helps us to observe things, and particularly to discriminate between similar things. We've long had machines that reliably and accurately take in data from the outside world. Think of autopilots that keep airplanes straight and level—or even land them when the weather is too bad to rely on the human pilot's own capacity for observation and discrimination. In fact, for virtually any task of observation (whether it involves vision or any of the four other senses), we now have the technology to build more sensitive, more accurate, and more reliable mechanical systems that can perform the task better than humans, for longer periods of time, under worse conditions, and with less susceptibility to illusion.

Consciousness helps us to plan. The ability to project far into the future and accurately predict the consequences of one's actions is often thought to be a mark of unusually high human intelligence, and one might be inclined to think that consciousness plays a role in this phenomenon. But there is no evidence to support this view. Machines routinely do tasks of this sort, ranging from the relatively simple projection made by my car's trip computer when it calculates where exactly on Interstate-5 I will run out of gas, to the Pentagon's enormously complex war simulation computers predicting in excruciating detail the outcome of a thermonuclear exchange between superpowers. Now consciousness may very well be required if one takes "to plan" to mean "to imagine experiencing future events," but this is hardly surprising since experiencing of any kind (even the imaginary sort) essentially involves qualia. But this sort of claim—that one must be capable of experiencing qualia in

order to experience qualia—is hardly illuminating, and there are other, qualia-free senses of planning that comport just as well with our standard use of the term.

Consciousness allows for memory. This is a much trickier claim to deal with. There seem to be three main types of memory, which I will dub *experiential* memory, *factual* memory, and *skill* memory. Experiential memories are just that: memories of experiences that the holder of the memory has been through. So when I think about my college graduation ceremony, I am calling up experiential memories. Factual memories could probably instead be called knowledge: when I remember that Jakarta is the capital of Indonesia, I am demonstrating knowledge by remembering a fact. Skill memories are memories of how to do something. My ability to ride a bicycle is a good example of a skill memory. Experiential and factual memories (and perhaps, some might argue, skill memories as well) are essentially qualitative in nature. This is why my recollection of watching the eruption of Mt. St. Helens counts as a full-fledged (experiential) memory, but a book filled with text and pictures chronicling the event does not. My memory has distinctly qualitative properties, while the book is just a series of pigment patterns on paper; it requires a conscious system to interpret those marks and to convert them into comprehensible semantic content. Or to use different terminology, my memory is intrinsically intentional, whereas the book's information has only derived intentionality. That is, the book relies on a bona fide intentional system (i.e., the person reading it) to "activate" it into something meaningful.[9]

But intrinsically intentional or qualitative memories are not interesting to us here for exactly the same reason that we rejected a particular definition of "to plan" above:

[9]The distinction between intrinsic and derived intentionality was touched on briefly in chapter three, or for a full review see Searle [95, ch. 3].

there is simply no question that they do require consciousness. A far more interesting issue is whether memory in a stripped down, qualia-less form requires consciousness. And here I think the answer is no. If we are content to call any record of a past fact or event a "memory"—and this would by necessity include memories with only derived intentionality—then we can see that the world is rife with memory-sustaining systems. Any non-fiction book (and perhaps any novel as well) is an example of this sort of memory, as is any photograph or audio recording. So as I see it, the claim that consciousness is a necessary condition for creating and recalling memories is either trivially false but not terribly relevant (for this answer relies on a very odd sense of the term "memory"), or else trivially true and utterly uninteresting (for *of course* all intrinsically qualitative phenomena involve qualia).

There is another possible link between these two phenomena that might be worth investigating. Is consciousness a sufficient condition for memory? Or to flip the relata: is memory a necessary condition for consciousness? This is harder to answer. On the one hand, it seems fairly clear that cases of anterograde or retrograde amensia indicate that one can be fully conscious even when missing some or all memories from before (retrograde) or after (anterograde) the onset of amnesia. But think of an extreme (and probably unrealistic) case in which no old memories are retained and no new memories are created (i.e., absolute retrograde amnesia accompanied by absolute anterograde amnesia). Though I do not wish to make the argument, I think there is real room to argue here that someone in this horrible state would not be conscious. Certainly he could continue to experience qualia, but the complete lack of connections between these qualia would render his life experiences very strange indeed. It is virtually impossible even to imagine what it would be like to be

in such a condition, and although I will tentatively claim that it would still qualify as a (devastatingly impoverished) form of consciousness, I readily concede that this issue is not at all clear cut.

Consciousness lets us learn. Perhaps the most frequent claim about the function of consciousness is that without it we would be doomed to repeat the same mistakes over and over again. But I think the capacity for learning cannot be the mark of consciousness, for any number of mechanical systems have demonstrated the ability to learn in some sense. These cases can be quite simple and straightforward, such as when a compact disc player learns my favorite Beatles songs when I program a play order for a particular disc, or when a nickel-cadmium rechargeable battery learns not to recharge fully in the future if it is not discharged completely the first few times. Other systems demonstrate more complex patterns of learning. Art Samuel's 1956 checkers program learned from both wins and losses, becoming more difficult to beat with every game that it played [89].[10] That learning seems not to require consciousness is not surprising given that learning is nothing more than the combination of a variety of other cognitive faculties, including perception, projection, memory, analysis, and perhaps others. As we have already seen, we have no reason to think that the first three require consciousness. I will not deal with the faculty of analysis here, mainly because it is too difficult to define the term into anything manageable.

I have listed four common functions or phenomena that are sometimes said to require consciousness, and I have claimed that all of these functions or phenomena can be performed by or found in unconscious systems such as robots, computers, or even perhaps

[10] As is the case with chess, the very best human checkers players can now be beaten by computers. *Chinook*, a program from the University of Alberta, first won the world checkers championship in 1998 [91, 90, 36].

books and photographs. But one might object that my claim is blatantly question-begging. If ostensibly non-conscious systems perform these tasks so well, shouldn't we at least entertain the possibility that they are conscious? This is the approach taken by Daniel Dennett, who has famously argued that if something behaves as if it is conscious, and if it helps us to think of that system as conscious, then there is every reason to think that it has consciousness that is just as authentic as yours or mine [25]. Or to put the complaint a slightly different way: is it fair for me to have shifted the burden away from myself (so that I no longer have to prove that these systems *lack* consciousness) and onto philosophers like Dennett (who now have to prove that they *have* consciousness)?

There is an easy but often overlooked answer to this charge. In fact, the answer seems so obvious and yet is so routinely ignored by otherwise intelligent philosophers that I am, to put it mildly, confused as to what I might be missing or misunderstanding about the response. But until I am shown why it is faulty, I will use it with confidence. The answer I have in mind is this: the claim that any system that *behaves* as if it were conscious actually *is* conscious is a claim based on the principle of behaviorism, plain and simple. Behaviorism says that all there is to the mind is its behavior, which is exactly the position that behavioral tests for consciousness rely on: if there is nothing to our mental life but behavior, then behavior should be an adequate test for consciousness or any other aspect of mental life. But behaviorism has long since been refuted as a solution to the mind/body problem. As I have already explained, it leaves something out of its account of the mind, and the thing that it leaves out is consciousness. To sum this point up: behavior is an adequate test only for, well, *behavior*; it is not a useful test for anything else. So it cannot

be used to reliably indicate the presence of consciousness.

Let me elaborate on these claims. As discussed earlier in this chapter, the central thesis of behaviorism is that mental states *just are* behavior, so mental states exist only insofar as behavior exists: feeling angry consists of nothing more than acting in a manner that others would interpret as being angry, and feeling pain is nothing over and above favoring the afflicted body part, looking for medical help, or somehow communicating to others the fact that one is in pain. I believe that behaviorism makes an interesting and helpful claim in pointing out that every mental state will affect the behavior of the holder of that state in *some* way, even if only in unusual and extreme conditions. For example, even an exceptionally strong stoic will behave in slightly different ways when feeling extreme pain than when healthy: he may be more curt with others, or his attention may be less focused, or he may find himself unable to remain stoic when additional sorts of pain are added—pain that he otherwise would be able to mask completely. The point is that mental states—even weaker and less vivid states than severe pain—will *somewhere* have an influence on the stoic's web of behavior.[11] But while behaviorism does make this positive contribution to our understanding of the mind, it fails to adequately account for a prominent feature of the mind. As I have argued throughout this dissertation, qualia are a definite and unignorable feature of our everyday experiences—in some sense they *are* experience!—and behaviorism leaves no room for this most basic and essential component of human mental life. Nowhere in its explanation of mental states does it address the issue of *what it's like* to be in a particular mental state, and this omission is what ultimately dooms the theory.

[11]This case is adapted from Hilary Putnam's [80] classic and to my mind utterly convincing "superspartan" counterexample to behaviorism.

Dennett is one of the best-known modern-day behaviorists, and he makes no bones about grounding much of his philosophical thought in the work of his one-time mentor Ryle. Other philosophers of mind are much less open about the (in my opinion deleterious) effect that behaviorism has had on shaping their views. Many of the standard positions in philosophy of mind, including eliminativism and functionalism, draw either directly or indirectly on behaviorist concepts. As I have already discussed, they suffer from the same problem that behaviorism does: they all leave qualia out of their accounts of the mind. And for this reason, I feel justified in passing the burden of proof on this issue to Dennett and his ilk. That is, I see no reason to think that appropriate behavior demonstrated by a biological or mechanical system is a sufficient test for consciousness in that system. Because behaviorism has failed as an explanation of mind/body interaction, the ball is now very solidly in the behaviorists' court, and until they offer an adequate argument for why we ought to be persuaded by intelligent or human-like behavior, I think we can comfortably insist on stricter standards for assessing the presence of consciousness. But I don't think this will happen. Behaviorism has failed in the past, and it seems very unlikely that a dusted-off, warmed-over version will work now.

Let's get back now to our project of trying to identify consciousness in other systems. If neither direct access nor behavioral tests will work, we might consider instead the underlying structural features of a system. Perhaps structural similarities between a particular system and a human brain might indicate the presence of consciousness in that system. But there are two closely related problems with this approach. First, it is not clear which level or levels of features are most relevant. In judging the presence of consciousness in a

computer, should we be convinced merely by high-level structural similarities between that computer and a mind (e.g., having roughly the same physical density and volume, or both being built on connectionist-flavored architectures), or should we demand similarities at lower physical levels as well? Answers to this question have of course been presupposed by some well-known solutions to the mind/body problem: token-token identity theory claims that similarities at the physical level are paramount, while functionalism asserts that only high-level structural features need be shared. But I have already discussed problems with these two theories, and we have seen that neither is specifically concerned with consciousness. It seems to me that the question of which level or levels of features are relevant cannot be answered without the possibility of direct access to other minds or reliable behavioral evidence of consciousness, which puts us right back where we started. Also—and this brings us to the second complaint one might have with using structural features as a means for determining the presence of consciousness—why do we think that structural features at *any* level are relevant? A common and devastating complaint made against type-type identity theory is that it is *chauvinistic*, in the sense that it seems arbitrary to declare that only humans or human-like systems are capable of consciousness, rational thought, or any other mental activity or phenomenon. Pick your favorite humanoid alien race and tell me that that they cannot possibly enjoy the same kind of consciousness that we do because their bodies are chromium- rather than carbon-based, or because their cerebral cortex is spread in a thin layer under the skin rather than nestled within the cranium, or because instead of neurons their brains are composed of a complex system of rubber bands and toothpicks. One could of course insist that this is the case, but it would be the height of

folly—and supremely racist—to do so. So it is not clear that conscious systems must have *any* structural features in common with human brains.

But now we're in deep trouble, for we've excluded every clear criterion by which to judge the presence of consciousness in other systems (even, disturbingly, in other human systems). This is true and yet it's puzzling, for we make exactly such judgments every day. Of course we can never be sure that these judgments are accurate, but this seems not to bother us a bit. After all, there are plenty of other things (causal relationships, for instance) that we know full well we can never be sure of, and yet we act as if they are no more epistemically suspect than, say, *a priori* mathematical truths. Does tobacco really cause cancer, or HIV cause AIDS? Does the impact of the cue ball really cause the motion of the eight ball? Does the external world exist at all? These issues do not trouble us, and neither does the question of whether consciousness is present in another human. But again, how are we able to make such judgments about consciousness so comfortably? I submit that it can only be through the use of a combination of the criteria outlined above. Under some circumstances a single criterion may suffice: a profoundly paralyzed but mentally lucid victim of so-called "locked-in" syndrome (caused by a lesion or stroke in the brain stem) may be thought of as conscious even when no consciousness-indicating behavior occurs, on the grounds that the structure of his cerebral cortex still appears to be sound. In other instances, failure of one criterion may trump success of the other: the behavior-derived illusion of consciousness presented by Joseph Weizenbaum's clever electronic "psychologist" program ELIZA is instantly broken once we understand that its responses are drawn from a vast database of prerecorded answers, using fairly simple logic

and straightforward syntactical analysis of questions and statements made by the human "patient" [115]. But in general we confirm the presence of consciousness by observing the right sorts of behavior and knowing that the behavior is supported by the right anatomical underpinnings. Once again, I admit that pure skepticism on this matter would be our safest bet, but that is no more realistic an option than is skepticism about whether the laws of gravity will hold tomorrow.

5.3 Consciousness and personhood

Some of the most interesting questions in philosophy lie in the branch of metaphysics that is concerned with personhood and personal identity. It is not obvious that the word "person" corresponds to any natural class of objects; we use the term in different ways at different times, with different purposes in mind. My project for this section of the chapter is to present a particular definition of personhood that I hope will highlight some of the practical implications of my view of consciousness. After giving my definition of the term, I will look specifically at how a qualia-based account of consciousness can influence how we think about personal identity, animals, machines, and death.

A quick but important warning is in order: this material will be the least rigorously argued and the most speculative portion of this dissertation. I freely admit this and am not troubled by it. Analytic philosophy is often accused of dwelling on irrelevant minutiae while avoiding Issues That Matter. Well, dealing with bigger issues requires grander, more sweeping, and perhaps less precise claims; our very biggest questions are probably better addressed through poetry or art than through strictly logical analytic philosophy. While I

hope to avoid shifting my approach *too* far in this direction, I hope the reader will keep in mind that the rest of this project will have a slightly different and perhaps looser feel than what has been presented so far.

5.3.1 What is a person?

The term "person" is most commonly used to pick out roughly the class of living human beings. This is not a perfect description of the everyday usage of the word, for different people would vary in their willingness to expand the term to include, say, intelligent aliens, or to restrict it to exclude brain-dead, severely brain damaged, permanently comatose, and perhaps criminally insane humans. But what seems relatively constant among different uses of the word is that its referents are all capable of experiencing qualia. That is, qualia-based consciousness seems to be a necessary condition on personhood. This, I trust, is a relatively uncontroversial claim. What is less clear is whether consciousness is also a sufficient condition for personhood. And here is where my use of the term "person" may depart from the norm. I want to claim that given the practical purposes that we have in mind when we typically employ the term, it is appropriate to consider consciousness a sufficient condition for personhood. That is, most of the questions we are trying to answer when we invoke the term "person" are questions that are concerned with all systems capable of experiencing qualia. This broad notion of personhood will perhaps become more comfortable as I work through some of the implications of including all conscious systems in our class of persons. Let's start with the issue of personal identity.

5.3.2 Personal identity

The philosophical problem of personal identity can be split into metaphysical and epistemic subproblems. First, we want to know *what makes it the case* that person a at time t_1 is identical to person b at time t_2, where t_1 and t_2 could be many years apart. Note that this is not the question of what properties persons a or b have that make them unique—we are not asking whether Ronald Reagan's identity is defined more by his years as an actor or by his two terms as president (one might call this concept "property identity"). Rather, we are wondering whether the person named Ronald Reagan who was a B-movie actor in the 1940s is numerically identical to the person named Ronald Reagan who served as president of the United States in the 1980s, and if so, what makes it the case that these two persons are identical (i.e., we're looking at what might be called "relational identity"). Second, we want to figure out *how we can know* when two people are the same over time. Because any answer to the epistemic question presupposes an answer to the metaphysical one, I will concentrate here only on the metaphysical issue.

The question of what makes two objects of any sort identical is a puzzling one. The question of what makes two *persons* identical is not only puzzling but also important in a way that identity of, say, pencils is not. It is important for several reasons. Perhaps most obviously, personal identity is crucial to our interpersonal relationships: we cannot maintain relationships based on friendship or love (or jealousy, or seething rage, or any other emotion) if we cannot identify the other participants in those relationships across time. It also plays a central role in the assigning of blame or responsibility. Courts cannot mete out punishment and society cannot issue praise unless the person who performed the act

that is being punished or rewarded can be identified at a later time. The notions of guilt or pride about past actions also make little sense without a practical way of tracking personal identity. It is pointless to plan for, anticipate, or fear future events that will directly affect you unless you can think of yourself as being identical to some particular person in the future. The list of reasons why personal identity is not just relevant but in fact crucial to our lives is a very long one.

Theories of personal identity are traditionally built either around mental criteria, physical criteria, or a combination of the two. For example, Anthony Quinton is a proponent of two mental criteria. He claims that two persons are identical if and only if they are both linked by a chain of memories and they share common character traits. Mental-based theories of this sort have a long history, ranging from Descartes (two persons are identical if they share the same spiritual soul or *res cogitans* [27]) to Derek Parfit (psychological continuity regulates both identity and survival, which he treats as two different concepts [72]). Theories of personal identity based on identity of the body are less popular, but also have prominent supporters such as Bernard Williams [117]. The theory that I want to propose falls more readily into the mental camp, but depending on what future research reveals about the production of consciousness, it may end up more properly described as a hybrid of the two views.

I claim that two persons are numerically identical if and only if they share the same consciousness. By this I mean not that they share memories or character traits (for I think these things can change radically as a person ages, without threatening the relation of identity that that person continues to bear to his younger self), but rather that their

qualia *register in the same place*. This is not a sophisticated or abstruse idea, but it is somewhat difficult to express. What I mean is this. Each person has a central (non-spatial) place where all of their qualia converge and are experienced. This place is not meant to correspond to any physical location, although it seems most natural to think of it as being somewhere within the confines of one's head. This has sometimes been called the "seat of consciousness" of a person, or a "Cartesian Theater." I dislike both of these terms: the former has distinctly religious overtones to my ear, while the latter suggests a well-known circularity problem involving homunculi who sit inside the theater and who presumably have Cartesian theaters within their own heads. I do not see my theory as presupposing any sort of spiritual soul, and homunculi play no role in my view of consciousness or personal identity. But these expressions might help to convey what I am driving at when I say that all of a person's qualia register in a certain place.

I realize that I have provided only a vague description of this concept of identical persons "sharing" a single consciousness, so let me try yet another way of expressing the idea. Consciousness is sometimes said to exist as a "unified field," which is meant to suggest that all of a person's concurrent qualia merge into a single unified experience, rather than registering on that person independently or in isolation from one another. For example, as I sit at my desk writing this, I am experiencing a number of different qualia: the smell of garlic coming from the kitchen, an itching in my nose caused by today's high pollen count, and the sound of a motorcycle outside, among others. But I do not experience these individually; rather, they coalesce into a single overall experience—they form a unified field of consciousness.[12] Now this phrase is generally used to describe the merging of the qualia

[12]I am told that Russell makes exactly this claim in his investigations into perception. It also arises, more

that one experiences *at a single moment*. But I think we can consider this phenomenon to occur not just *synchronically*, but also *diachronically*. That is, it makes just as much sense to think of our qualia being unified *across* time as it does to think of them being unified *at a single* time. And this notion of qualia being unified across time is part of what I am driving at when I say that personal identity consists in the sharing of a single consciousness. If person a at time t_1 is numerically identical to person b at time t_2, then a's field of consciousness at t_1 must be unified (across time) with b's field of consciousness at t_2. So we now have at our disposal a number of different metaphors for understanding my claim about the necessary and sufficient conditions for personal identity. Even if no single metaphor is quite precise enough to fully capture my meaning, I hope that some combination of them will allow the reader to triangulate on the concept that I am trying to convey.

Theses similar to mine have been advanced in many places in the philosophical literature. In arguing for his version of the memory criterion for personal identity, Quinton refers to the notion of a person's qualia collecting in or registering on a particular seat of consciousness or mind: "Two soul-phases belong to the same soul ... if they are connected by a continuous character and memory path. A soul-phase is a set of contemporaneous mental states *belonging to the same momentary consciousness*." [emphasis mine] [81, p. 59] Thomas Reid also employs a concept similar to mine when he talks about persisting subjects of thought:

> I am not thought, I am not action, I am not feeling; I am something

recently, in Searle: "[C]onscious experiences, such as the taste of beer or the smell of a rose, always come as part of a unified conscious field. I do not, for example, right now feel just the pressure of the shirt on my back and the aftertaste of coffee in my mouth, and the sight of the computer screen in front of me; I have all of those as part of a single unified conscious field." [98, p. 271]

> that thinks and acts and suffers. My thoughts, and actions, and feelings, change every moment; they have no continued, but [rather] a successive, existence; but that *self*, or *I*, to which they belong, is permanent, and has the same relation to all the succeeding thoughts, actions, and feelings which I call mine. [emphasis his] [82, p. 109]
>
> The operations of our minds are all successive, and have no continued existence. But the thinking thing has a continued existence, and we have an invincible belief, that it remains the same when all its thoughts and operations change. [82, p. 111]

While Reid is gesturing in the right general direction, I am not entirely comfortable with his description of a *thing* that thinks. This smacks too much of Cartesian dualism for my taste. My claim is subtly different: I want to say there is some property that all of my qualia have in common and that your qualia lack, and that there is an entirely different property that your qualia have in common but that mine lack. I do not want to make a metaphysical claim about the existence of a thinking *substance*, whether it be spiritual or material.

An example may help to clarify my view further. Imagine two people, Alice and Bill, sitting in the stands of a baseball game. During the game, they both see the same green field and white uniforms, hear the same cheers from the crowd, and smell the same greasy stadium food. The qualia they experience during those three hours are likely very similar—we can even imagine that they are type-identical without running into any logical problems. But there are two distinct sets of qualia involved here: there are Alice's qualia, and then there are Bill's. And each set of qualia, I claim, converges on and registers within a different mental "place." Or to use different terminology, each is unified (both diachronically and synchronically) with a different set of other qualia. Bill's qualia may be *type*-identical to Alice's, but they cannot be *token*-identical because they are, for lack of a better way to

put it, his qualia and not hers. He is the subject of his qualia, and she is the subject of hers. There is never any ambiguity about who it is that experiences a particular quale, and there are no free-floating, unattached, subject-less qualia.[13] This is an obvious but very important point.

One might be tempted to respond to this by saying that each person's qualia are lumped together in virtue of memory links not only between the mental states that have those qualia, but also to past events in that person's life. That is, even though the qualia experienced by Alice and Bill are type-identical, we can lasso all of Alice's baseball-game qualia by seeing that the mental states with those qualia all are linked (either directly or through a chain of other mental states) to her memories of learning to swim as a child, her first day of college, and skydiving on her fortieth birthday. Similarly, Bill's baseball-game qualia might be grouped together in virtue of being linked to his memories from *his* life. Or to put this criticism another way: perhaps my view is just a recasting of the old memory-based theory of personal identity, whereby two persons are identical either if they share common memories, or if one has memories of having been the other.

But this criticism misunderstands my claim. I don't see how memory can have anything to do with personal identity. Imagine that Alice and Bill attend the baseball game while under the influence of powerful amnesia-inducing drugs: not only do they remember nothing from their previous lives, but they are unable to lay down any short- or long-term memories of the qualia they experience while watching the game. Although we would

[13]It is for this reason that when a tree falls in the forest with no one around to hear it, it makes no sound. The sound of a tree falling is, properly, a quale (for compression waves in air are not converted into sound until they register as such on a conscious subject), and if there are no subjects around to have that quale, then there can in fact be no quale present. This reveals frustratingly little about the sound of one hand clapping, however.

consider their lives for those three hours to be strange and impoverished, I can't imagine anyone denying that they are still persons, and that they continue to be identical to pre-drug Alice and pre-drug Bill, and that they are identical to the (not yet existing) post-drug Alice and post-drug Bill. And the reason we have these intuitions is that before, during, and after the influence of this drug, qualia produced by Alice's brain all arrive in one place (i.e., they are unified), and qualia produced by Bill's brain all arrive in a different place (i.e., are unified in an entirely different field). Even if someone were to have no memories at any point in his life, there would be some property of his qualia that would remain constant over the course of that life; there would be some property that all of his qualia would share. That property is the property of registering in a particular *mental place*, or of being unified in a particular field. And that is what distinguishes my mental life from yours, rather than memory or character or values or attitudes.[14]

I will try one more tack to explain this point. Let's say that Alice is a normal person leading a normal life, and that Bill is an equally normal person whose qualia all happen to be type-identical to qualia experienced by Alice. Imagine that this has been the case throughout their lives. Maybe this is due to the machinations of a mad scientist

[14]The best known argument against my position is probably by Hume, who insists that when he turns his gaze inward, there is no "self" apparent—he encounters only a bundle of perceptions: "For my part, when I enter most intimately into what I call *myself*, I always stumble on some particular perception or other, of heat or cold, light or shade, love or hatred, pain or pleasure. I never can catch *myself* at any time without a perception, and never can observe any thing but the perception." [47, p. 252] But he later expresses doubts about this claim when he notes that he has failed to account for why bundles of perceptions are made up of those particular perceptions and not others: "[W]hen I proceed to explain the principle of connexion, which binds [our perceptions] together, and makes us attribute to them a real simplicity and identity; I am sensible, that my account is very defective.... [A]ll my hopes vanish, when I come to the principles, that unite our successive perceptions in our thought or consciousness. I cannot discover any theory, which gives me satisfaction on this head." [47, pp. 635–6] He offers no explanation for how this bundling occurs, and neither can I: I do not know why my qualia register in my mind and not in someone else's. But since he realizes that this is a real question that needs to be answered—that more needs to be said about persons than that they are just collections of perceptions—I am not overly troubled by his original complaint.

who has rigidly controlled the environments in which they've lived, or maybe it is just coincidence. Because all of their qualia have been type-identical, their memories are also all type-identical. Now according to the memory criterion, we either have to say that they are numerically identical persons (which I trust most people will find an unsatisfying description of the situation), or else we have to group some of these past and present qualia together and call them Alice's, and group the rest together and call them Bill's. But the memory criterion offers us no guidance on how to decide whether to assign a particular quale to Alice or to Bill; the memory criterion suggests that it would be perfectly legitimate to lump together some of the qualia from Bill's body and some of the qualia from Alice's body into a single person. And that seems crazy. I consider this reason enough to look for alternatives to the memory criterion. The bodily criterion offers the advantage of keeping the mental states produced by Alice's brain and the mental state produced by Bill's brain distinct, but it faces problems of its own. To state just one: our bodies change routinely and even drastically without threatening personal identity. And even if one is committed to the bodily criterion, we can adapt the example by changing Bill's sex and giving him a body type-identical to Alice's, in which case we face the same problem that caused trouble for the memory criterion. So if we can't rely on memory and we can't rely on bodily identity to distinguish the two people and their qualia, then what is it about Alice and Bill that justifies our intuition that there is one group of qualia that is appropriate to call Alice's, and another non-overlapping group that is appropriate to call Bill's? I submit that her qualia all have the property of registering on *her* mind—in her seat of consciousness, or her Cartesian Theater, or her unified field of consciousness—and not on Bill's, and that his have

the importantly different property of registering on *his* mind. This of course means that it is impossible for two numerically distinct people to have completely type-identical qualia (for they would have to differ with respect to this property), but that has no important implications that I know of. So to sum up: my solution to the problem of personal identity says that two persons are identical if and only if they have the same seat of consciousness.[15]

Locke is standardly interpreted as being a proponent of the memory criterion. But I think there is an alternative way to read him that supports my view. Let's look at a key passage that is often taken to indicate his belief in the memory criterion:

> [S]ince consciousness always accompanies thinking, and it is that which makes every one to be what he calls self, and thereby distinguishes himself from all other thinking things: in this alone consists personal identity, i.e., the sameness of a rational being; and as far as this consciousness can be extended backwards to any past action or thought, so far reaches the identity of that person; it is the same self now it was then; and it is by the same self with this present one that now reflects on it, that that action was done. [59, pp. 39–40]

Locke's mention of extending consciousness backwards is generally taken to mean having memories of past events, so he is seen as claiming that memory is the sole determiner of personal identity. But it is perfectly consistent with the rest of his writings on personal identity to read him as talking about a seat of consciousness rather than about memory—so for person y to extend his consciousness back to an action a performed by person x would mean not that person y remembers having done action a, but rather that the seat of consciousness of person y is the same as the seat of consciousness of the person x who performed action a.

[15]For what it's worth, this suggests a monadic as opposed to constructivist concept of persons; I do not know how to make sense of the idea of splitting or duplicating a seat of consciousness, such as is discussed by Parfit [72] and Nagel [68].

Interpreting Locke in this way allows him to avoid certain problems that face memory-based theories, problems that are so basic and so apparent that it would be truly strange for him to have proposed a solution that is susceptible to them. One problem involves amnesia. On the memory criterion, severe lapses in memory would seem to destroy the person whose memories are erased while creating an entirely new person in their body. A second problem is that a shift in memories from one body to another (say, one person forgets having performed some action while a second person has a type-identical memory implanted through neurosurgery) would mean that the person inhabiting the first body would actually move to the second body. We could call these the "one body/several persons problem" and the "one person/several bodies problem" respectively. A third difficulty is the "brave officer objection" given by Reid [83]. Imagine a man at three stages in his life: person x is a child who is flogged for some misbehavior, person y is a military officer who captures an enemy standard, and person z is a retired general reminiscing about his life. If person z remembers taking the standard as y, and y remembers being flogged as x, but old age prevents z from remembering being flogged as x, then by the memory criterion z *is not* identical to x, but by transitivity of the identity relation, z *is* identical to x. This contradiction argues against the memory criterion, and, I would think, also argues against reading Locke as proposing that criterion. Using the "seat or unified field of consciousness" criterion solves both of these problems: if x, y, and z all share the same seat or unified field of consciousness, then they are numerically identical persons. If they don't, then they aren't. And on this reading of Locke, amnesia and memory shifts cease to be troubling issues. As long as the locus of consciousness remains the same, one person remains safely ensconced in one body no matter

how many memories are wiped out or transferred in.

Although I think the reading of Locke that I have presented is a perfectly natural one that happens to strengthen his theory of personal identity significantly, I should reiterate that this is a nonstandard interpretation of his theory. Because of this, I want to present and discuss one more quotation from Locke to bolster my claim:

> For as far as any intelligent being can repeat the idea of any past action with the same consciousness it had of it at first, and with the same consciousness it has of any present action; so far it is the same personal self. For it is by the same consciousness it has of its present thoughts and actions, that it is self to itself now, and so will be the same self, as far as the same consciousness can extend to actions past or to come.... [59, pp. 40–41]

Two phrases in this passage are confusing on the standard reading but are quite clear on my interpretation. If Locke really means to refer to memory, then we might wonder why he describes this simple concept by saying that one "repeats an idea of a past action with the same consciousness one had of it at first." Why does he not just say that two persons are identical if they share memories, or if the later one has memories of having been the earlier one? On the other hand, if he is claiming that the seat of consciousness is the same in any two identical persons, then the expression he uses is appropriate: two people with the same seat of consciousness will indeed experience actions "with the same consciousness." Second, his mention of consciousness "extending to actions to come" is odd. If by this he means that in the future I will remember things about the present moment, then he should have spoken of future consciousness extending backwards, rather than speaking of present consciousness extending forwards. This is a minor point that I do not mean to rely on too heavily, but once again it does seem more natural to understand this expression in terms of a constant place where qualia register: my future qualia will converge on the same seat of

consciousness that my present qualia converge on.

I want to end this discussion by pointing out two important ways in which the issue of personal identity differs from the issue of identity of objects other than persons. First, personal identity is always a one-to-one relation: any given person at time t_1 can be identical to at most one person at time t_2. We simply don't know what it would mean to say that someone is identical to two (non-identical) people at a given time in the future or past. To put this another way, most of us have the firm intuition that persons are indivisible. Second, we also have the intuition that questions about personal identity always have determinate answers. If I ask whether Nick and Nicholas are the same person, there is no room for interpreting either the question or the facts about the world that provide an answer to the question. They simply *are* or *are not* numerically identical persons. We may not be able to tell if they are identical in a particular case, but this is an epistemic rather than a metaphysical issue. Reid realizes at least the second of these two features of personal identity:

> The identity, therefore, which we ascribe to bodies, whether natural or artificial, is not perfect identity; it is rather something which, for the conveniency of speech, we call identity. It admits of a great change of the subject, providing the change be gradual; sometimes, even of a total change.... It has no fixed nature when applied to bodies; and questions about the identity of a body are very often questions about words. But identity, when applied to persons, has no ambiguity, and admits not of degrees, or of more and less. It is the foundation of all rights and obligations, and of all accountableness; and the notion of it is fixed and precise. [82, p. 112]

These two characteristics of personal identity can break down where identity of non-personal objects are concerned. If I break two pencils a and b in half and then glue the bottom half of a to the top half of b, and the bottom half of b to the top half of a, there

is a sense in which we want to say that the original pencils a and b are each identical to both of the resulting pencils, showing that identity of pencils may not be strictly one-to-one. And even if we don't want to make that claim, we certainly do want to admit that there is an ambiguity to the situation that prevents us confidently stating which pencils are identical to which. Or consider an example given by Parfit, in which a watch that is sent in for repair sits in a disassembled heap for a week and then is reassembled [73, p. 203]. We might very well express doubts as to whether the reassembled watch is numerically identical to the original broken watch, and we would certainly hesitate to say that the original watch is numerically identical to the pile of disassembled watch parts. Now both of these features are explainable in the case of non-personal objects. The possibility of pencil identity not being one-to-one stems from the lack of essential features of pencils—there is no core notion of "pencilhood," so there is no straightforward description of what makes two things the same pencil. And ambiguity in the watch case arises from our common reliance on spatiotemporal contiguity as a good indicator of identity. This contiguity is disrupted when the watch is disassembled, and we are left wondering whether the disruption is severe enough to break the identity relation. The fact that personal identity does not suffer from these sorts of ambiguity indicates that there is an essential feature of personhood, and that that feature is not spatiotemporal contiguity. Given the vagueness of many memories, that feature is probably not memory related either. But saying that personhood consists in experiencing qualia that all register in a particular seat or unified field of consciousness, and that two persons are identical if and only if they share the same seat or unified field of consciousness allows us to account for the unambiguous, one-to-one nature of the personal

identity relation.

To sum up: I have proposed a definition of personal identity that is built on a qualia-based concept of consciousness; I have shown how we can interpret philosophers such as Quinton, Reid, and Locke in such a way as to draw support for this view; and I have sketched out how and why the concept of personal identity differs from our concept of non-personal identity. In the final sections of this dissertation I will very briefly discuss some implications of assigning the status of "person" to any system that experiences qualia.

5.3.3 Animal rights

It is extremely likely that at least some animals qualify as persons under my definition of personhood. I don't know how far down the phylogenetic scale qualia and personhood extend, and I haven't a clue how to conclusively answer this question—or indeed whether there even is a determinate answer. Donald Griffin, the co-discoverer of echolocation in bats in the 1950s, is generally thought to have founded the modern study of animal consciousness, or cognitive ethology.[16] He took ethological evidence to point to consciousness in a wide variety of animals, thus starting a tradition that persists today, virtually unquestioned. Witness the signs of animal consciousness cited in a recent article by the *New York Times*: "[I]n Arizona an African Gray parrot named Alex can identify colors and shapes as well as any preschooler. In Georgia a bonobo ape named Kanzi converses with his trainer via computer keyboard and watches Tarzan movies on television.... Animal enrichment programs featuring mental puzzles disguised as toys and treats have become a standard part of daily life at zoos." [31]

[16] See [38] for some of Griffin's early claims, and [39] for more refined arguments.

But we should not be too quick to accept Griffin's claims. As I have argued in several places in this dissertation, behavior alone does not make for an adequate test. It is simply too easy to be lured into mistakenly ascribing a full mental life to a system that has been cleverly programmed. Surprisingly, even a cloaked behaviorist like Dennett agrees on this point: "It is in fact ridiculously easy to induce powerful intuitions of not just sentience but full-blown consciousness (ripe with malevolence or curiosity or friendship) by exposing people to quite simple robots made to move in familiar mammalian ways at mammalian speeds." [26] Given the danger of relying on behavior alone, a better bet (and probably the best we can do) is to rely on some blend of behavioral and neurophysiological evidence. Have you got sophisticated social behavior, some capacity to learn, and a relatively large cerebral cortex? You probably qualify. But if you're a fruit fly with utterly rigid mating behavior, non-adaptive movement toward or away from particular chemicals, and a minimally complex brain substructure, then we're probably justified in excluding you from the club.[17] Peter Singer [100] and Colin McGinn [64], to name two of the most respected philosophers who have worked on this issue, both claim that animal consciousness exists right down to the level of insects. While many people would want to place the bar a bit higher than this, I think virtually no one would be willing to take seriously the Cartesian position that *all* non-human animals are simply automata with appropriate (but purely mechanical) responses to ersatz pain, hunger, etc.

So what are we to do with my claim that many animals are conscious and hence

[17]Judging physical similarity between species is perhaps harder than one might think, given that the recently completed mapping of the human genome has revealed humans to be *much* closer genetically to other animals than we had previously guessed: we now know humans to have roughly 30,000 genes instead of the 100,000 we expected to find, as compared to 15,000 genes in fruit flies and 19,000 in nematodes. It is now speculated that the complexity of the human phenotype comes not from the sheer number of our genes but rather from the unexpectedly large number of proteins that each gene can produce.

are in some sense persons? While I will offer no concrete prescriptions for behavioral or legal changes, I can suggest that as persons, animals may have certain rights that are currently denied to them. I don't want to recommend that we honor any rights in particular, for I recognize that many different sorts of considerations should come into play when humans make decisions about the lives—and, frequently, deaths—of other species: intelligence, sophistication of thought processes, population size and growth rate, capacity for emotion, likelihood of free will, and perhaps language usage. For example, it is probably not appropriate to say that all animals that can experience qualia should enjoy the right of free assembly or the right to trial by jury. But I do think we should think very carefully about whether there are any animal rights that we should take more seriously. A very small sampling of questions that seem appropriate to consider include the following:

- Should we kill conscious animals for food, fashion, or furniture, especially when non-conscious alternatives are available? If so, which animals, and under which circumstances?
- Should we use them as a source of labor? If so, under what conditions is this appropriate?
- Is there a particular standard of living that we are obligated to provide for animals under our care?
- Should we test cosmetic, pharmaceutical, or food products on them?
- Is it right to use them for human entertainment, especially when this is to the detriment to the health or happiness of the animal?

- Are there any animal training techniques that are inappropriate?

5.3.4 Machine rights

Chapter four was dedicated to arguing that no current (non-biological) machine is likely to be conscious, and that this will remain the case until we are able to identify and replicate the causal mechanisms that produce consciousness (i.e., until we solve the "hard problem" referred to in chapter one). But if neurobiology, computer science, psychology, and philosophy continue to make strides at the pace we've seen over the last twenty years or so, these problems may one day be solved and we may find ourselves—for better or worse—ushering in the era of conscious machines. It would behoove us to consider the implications of this development before it occurs.

Many of the questions we asked regarding animals should also be asked about machines, since conscious machines would presumably play roles similar to those played by animals today. But new, machine-specific questions arise as well. To wit:

- Is it morally acceptable to turn off or destroy a conscious machine? If John McCarthy's famous thermostat [63] (or, more likely, some future version of it) really does have three qualitative states (the thought that it's too hot in here, the thought that it's too cold in here, and the thought that it's just right in here), then would it ever be ethical to replace or destroy this thermostat? To push the question to an absurd extreme: will we eventually be obligated to construct special "preserves" where obsolete but still qualia-capable machines go to experience their qualia in peace?

- Conscious machines would presumably be used to relieve us of onerous or unpleasant chores (think of modern-day automobile assembly line robots). This raises the issue of whether they are due certain privileges that we grant to the human workers that they would replace. Should we give these machines vacation days? How about the right to unionize?

- Would it be appropriate to reward conscious machines in some way when they do their tasks well? If so, how?

- Do machines have responsibilities to act in certain ways? We sometimes deem it appropriate to punish the inventor of a machine that harms other people or society (e.g., computer hackers go to jail if their viruses damage or disrupt other machines). Would it ever be right to punish a machine that acts improperly? What would punishment consist of?

Again, I cannot answer these questions here. I cannot even suggest ways in which we might begin to look for answers. I simply want to emphasize the importance both of asking these questions, and of living in a manner that is consonant with whatever answers we do arrive at. If my analyses of consciousness and personhood are at all accurate, then we cannot morally do otherwise.

5.3.5 Death

I end this dissertation on a very brief, but I hope positive, note. I wish to suggest that it is conceivable—though extremely unlikely—that consciousness, and hence personhood, could continue after one's body has died. If the right sorts of qualia continue to

occur, one might be able to "live" in some sense even after the death of one's body. Not just any qualia will do; the qualia that I experience today do not make it the case that my late grandmother continues to exist as a person today. Rather, qualia must continue to register in the right mental place—in the right unified field of consciousness—if one's mental life is to continue. Specifically, they must register in the same seat of consciousness on which qualia converged during that person's embodied life. This eventuality is not ruled out by the complete destruction of the body after death (in which case the bodily criterion for personal identity would suggest that a person could not survive), or even if all memories and character traits are erased or changed (which would cause the various memory criteria to deny the possibility of an afterlife). Now I want to be clear that I can offer no explanation of how this might be possible, and in fact I very much doubt that it *is* possible. But until we know much more about how qualia are produced and why a particular quale has as its subject a particular person and not any other person, we cannot cateogrically rule it out. We *can* say with certainty that in order for me to survive the death of my body, there would have to be a way for the right sorts of qualia to occur without any processes occurring in my (dead or destroyed) brain. That is, qualia must not be supervenient on the physical realm. Again, this scenario seems highly unlikely, but according to my analyses of consciousness and personhood it is not logically impossible.

Of course, whether mental immortality is a *good* thing is an entirely different question. I will leave this to the reader to decide for himself, but I happen to be entirely in favor of it.

Bibliography

[1] Bryan Adams, Cynthia Breazeal, Rodney A. Brooks, and Brian Scasellati. Humanoid Robots: A New Kind of Tool. IEEE *Intelligent Systems*, 14(4):25–31, July/August 2000.

[2] David M. Armstrong. What is Consciousness?, in *The Nature of Mind*, pp. 55–67. Cornell University Press, Ithaca, 1981.

[3] David M. Armstrong. *Universals: an Opinionated Introduction*. Westview Press, Boulder, Colo., 1989.

[4] David M. Armstrong. *A Materialist Theory of the Mind*. Routledge, London, 1993.

[5] Bernard J. Baars. Contrastive Phenomenology: A Thoroughly Empirical Approach to Consciousness. *Psyche*, 1, 1994. Available only at http://psyche.cs.monash.edu.au /v2/psyche-1-6-baars.html.

[6] Bernard J. Baars. *In the Theater of Consciousness: the Workspace of the Mind*. Oxford University Press, New York, 1997.

[7] Ned Block. Troubles with Functionalism, in Ned Block, ed., *Readings in Philosophical Psychology*, vol. 1. Harvard University Press, Cambridge, Mass., 1980.

[8] Ned Block. Inverted Earth, in James E. Tomberlin, ed., *Philosophical Perspectives*, vol. 4, pp. 52–79. Ridgeview, Atascadero, Calif., 1990.

[9] Ned Block. On a Confusion About a Function of Consciousness. *Behavioral and Brain Sciences*, 18:227–247, 1995.

[10] Franz Brentano. *Psychology from an Empirical Standpoint*. Humanities Press, New York, 1973. A. Rancurello, D. B. Terrell, L. McAlister, trans.

[11] Rodney A. Brooks. Intelligence without Representation, in John Haugeland, ed., *Mind Design II*, pp. 395–420. M.I.T. Press, Cambridge, Mass., 1997.

[12] Rodney A. Brooks, Cynthia Breazeal, Matthew Marjanović, Brian Scassellati, and Matthew M. Williamson. The Cog Project: Building a Humanoid Robot, in C. Nehaniv, ed., *Computation for Metaphors, Analogy and Agents*, vol. 1562 of *Springer Lecture Notes in Artificial Intelligence*, pp. 52–87. Springer-Verlag, Berlin, 1999.

[13] Tyler Burge. Two Kinds of Consciousness, in Ned Block, Owen Flanagan, and Güven Güzeldere, eds., *The Nature of Consciousness*, pp. 427–434. M.I.T. Press, Cambridge, Mass., 1997.

[14] Peter Carruthers. Brute Experience. *Journal of Philosophy*, 86(5):258–269, 1989.

[15] David Chalmers. *The Conscious Mind*. Oxford University Press, Oxford, 1996.

[16] David Chalmers. Facing Up to the Problem of Consciousness, in Stuart R. Hameroff, Alfred W. Kaszniak, and Alwyn C. Scott, eds., *Toward a Science of Consciousness: The First Tucson Discussions and Debates*, pp. 5–28. M.I.T. Press, Cambridge, Mass., 1996.

[17] Jennifer Church. Fallacies or Analyses? *Behavioral and Brain Sciences*, 18:251–252, 1995.

[18] Paul M. Churchland. Reduction, Qualia, and the Direct Introspection of Brain States. *Journal of Philosophy*, 82(1), 1985.

[19] Paul M. Churchland. *Matter and Consciousness.* M.I.T. Press, Cambridge, Mass., 1988.

[20] Paul M. Churchland. *The Engine of Reason, the Seat of the Soul: a Philosophical Journey into the Brain.* M.I.T. Press, Cambridge, Mass., 1995.

[21] Francis Crick. *The Astonishing Hypothesis.* Simon and Schuster, New York, 1994.

[22] Francis Crick and C. Asanuma. Certain Aspects of the Anatomy and Physiology of the Cerebral Cortex, in David E. Rumelhart and James L. McClelland, eds., *Parallel Distributed Processing*, vol. 2, pp. 333–371. M.I.T. Press, Cambridge, Mass., 1986.

[23] Francis Crick and Christof Koch. Towards a Neurobiological Theory of Consciousness, in *Seminars in the Neurosciences*, vol. 2, pp. 263–275. Saunders Scientific Publications, Philadelphia, 1990.

[24] Daniel Dennett. Quining Qualia, in Anthony J. Marcel and Edoardo Bisiach, eds., *Consciousness in Contemporary Science*. Oxford University Press, New York, 1988.

[25] Daniel Dennett. *Consciousness Explained*. Little, Brown, & Co., Boston, 1991.

[26] Daniel Dennett. Cog: Steps Towards Consciousness in Robots, in Thomas Metzinger, ed., *Conscious Experience*. Ferdinand Schoningh, Paderborn, 1995.

[27] René Descartes. Meditations on First Philosophy, in John Cottingham, Robert Stoofhoff, and Dugold Murdoch, eds., *The Philosophical Writings of Descartes*, vol. 2, pp. 1–62. Cambridge University Press, Cambridge, 1641, 1984.

[28] Fred Dretske. *Seeing and Knowing*. University of Chicago Press, Chicago, 1969.

[29] Fred Dretske. Conscious Experience. *Mind*, 102(406):263–283, April 1993.

[30] Fred Dretske. Phenomenal Externalism, in Enrique Villanueva, ed., *Perception*. Ridgeview, Atascadero, Calif., 1996.

[31] Emily Eakin. No Longer Alone: The Scientist Who Dared to Say Animals Think. *New York Times*, CL(51653):A17, February 3, 2001.

[32] Gerald Edelman. *Bright Air, Brilliant Fire: On the Matter of the Mind*. Basic Books, New York, 1992.

[33] Owen Flanagan. *Consciousness Reconsidered*. M.I.T. Press, Cambridge, Mass., 1992.

[34] Jerry A. Fodor and Zenon W. Pylyshyn. Connectionism and Cognitive Architecture. *Cognition*, 28:3–71, 1988.

[35] Sigmund Freud and trans. by Joan Rivere. The Unconscious, in *Collected Papers*, vol. IV. Hogarth, London, 1953.

[36] Matthew L. Ginsberg. Computers, Games and the Real World. *Scientific American*, November 1998.

[37] Alvin Goldman. Consciousness, Folk Psychology, and Cognitive Science. *Consciousness and Cognition*, 2(4):364–382, 1993.

[38] Donald R. Griffin. *The Question of Animal Awareness: Evolutionary Continuity of Mental Experience*. Rockefeller University Press, New York, 1976.

[39] Donald R. Griffin. *Animal Minds*. University of Chicago Press, Chicago, 1992.

[40] Güven Güzeldere. Is Consciousness the Perception of What Passes in One's Own Mind?, in Thomas Metzinger, ed., *Conscious Experience*, pp. 335–357. Ferdinand Schoningh, Paderborn, 1995.

[41] Güven Güzeldere. The Many Faces of Consciousness: A Field Guide, in Ned Block, Owen Flanagan, and Güven Güzeldere, eds., *The Nature of Consciousness*, pp. 1–67. M.I.T. Press, Cambridge, Mass., 1997.

[42] Stuart R. Hameroff. *Ultimate Computing: Biomolecular Consciousness and Nanotechnology*. North-Holland, New York, 1987.

[43] Stuart R. Hameroff, S. Rasmussen, and B. Mansson. Molecular Automata in Microtubules: Basic Computational Logic of the Living State?, in C. Langton, ed., *Artificial Life*, S.F.I. Studies in the Sciences of Complexity. Addison-Wesley, New York, 1988.

[44] Gilbert Harman. The Intrinsic Quality of Experience, in James E. Tomberlin, ed., *Philsophical Perspectives*, vol. 4, pp. 31–52. Ridgeview, Atascadero, Calif., 1990.

[45] Gilbert Harman. Explaining Objective Color in Terms of Subjective Experience, in Enrique Villanueve, ed., *Perception*. Ridgeview, Atascadero, Calif., 1996.

[46] Carl G. Hempel. The Logical Analysis of Psychology, in Ned Block, ed., *Readings in Philosophy of Psychology*, vol. 1. Harvard University Press, Cambridge, Mass., 1980.

[47] David Hume. *A Treatise of Human Nature*. Oxford University Press, Oxford, 2nd edition, 1739, 1978. Peter H. Nidditch, ed.

[48] Thomas H. Huxley. *Lessons in Elementary Physiology*. Macmillan, New York, 1923.

[49] Ray Jackendoff. *Consciousness and the Computational Mind*. M.I.T. Press, Cambridge, Mass., 1987.

[50] William James. *Psychology: The Briefer Course*. University of Notre Dame Press, Notre Dame, Indiana, 1892, 1985. Gordon Allport, ed.

[51] William James. *Principles of Psychology*, vol. 1. Dover, New York, 1950.

[52] John F. Kihlstrom. The Continuum of Consciousness. *Consciousness and Cognition*, 2(4):334–354, 1993.

[53] Stephen LaBerge and P. G. Zimbardo. Smooth Tracking Eye-Movements Discriminate Both Dreaming and Perception from Imagination, in *Towards a Science of Consciousness*, Tucson, Ariz., April 2000.

[54] Karl S. Lashley. Cerebral Organization and Behavior, in Harry C. Solomon, Stanley Cobb, and Wilder Penfield, eds., *The Brain and Human Behavior*. Williams and Wilkins, Baltimore, 1958.

[55] Gottfried Wilhelm Leibniz. Discourse on Metaphysics [abridged], in G. H. R. Parkinson, ed., *Philosophical Writings*. Dent, London, 1686, 1934.

[56] Gottfried Wilhelm Leibniz. On the Principle of Indiscernibles, in G. H. R. Parkinson and Mary Morris, eds., *Philosophical Writings [of] Leibniz*. Dent, London, 1696, 1973.

[57] Brian Loar. Phenomenal States, in James E. Tomberlin, ed., *Philosophical Perspectives*, vol. 4, pp. 81–108. Ridgeview, Atascadero, Calif., 1990.

[58] John Locke. *An Essay Concerning Human Understanding*. Clarendon, Oxford, 1690, 1975. Peter H. Nidditch, ed.

[59] John Locke. Of Identity and Diversity, in John Perry, ed., *Personal Identity*, page 39. University of California Press, Berkeley, 1975.

[60] William Lycan. *Consciousness*. M.I.T. Press, Cambridge, Mass., 1987.

[61] William Lycan. Consciousness as Internal Monitoring, in James E. Tomberlin, ed., *Philosophical Perspectives*, vol. 9, pp. 1–14. Ridgeview, Atascadero, Calif., 1995.

[62] David Marr. *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*. W. H. Freeman, San Francisco, 1982.

[63] John McCarthy. Ascribing Mental Qualities to Machines, in Martin Ringle, ed.,

Philosophical Perspectives in Artificial Intelligence, vol. 7 of *Harvester Studies in Cognitive Science*. Harvester Press, Brighton, Sussex, 1979.

[64] Colin McGinn. Animal Minds, Animal Morality. *Social Research*, 62(3), 1995.

[65] Colin McGinn. *The Mysterious Flame: Conscious Minds in a Material World*. Basic Books, New York, 1999.

[66] Brian P. McLaughlin. Philosophy of Mind, in Robert Audi, ed., *The Cambridge Dictionary of Philosophy*, pp. 597–606. Cambridge University Press, Cambridge, 1995.

[67] Marvin Minsky. *The Society of Mind*. Simon & Schuster, New York, 1985.

[68] Thomas Nagel. Brain Bisection and the Unity of Consciousness. *Synthese*, 22, 1971.

[69] Thomas Nagel. What Is It Like to Be a Bat?, in *Mortal Questions*. Cambridge University Press, Cambridge, 1979.

[70] Allan Newell and Herbert A. Simon. Computer Science as Empirical Inquiry: Symbols and Search, in John Haugeland, ed., *Mind Design*. M.I.T. Press, Cambridge, Mass., 1981.

[71] Bente Pakkenberg and Hans Joergen Gundersen. Neocortical Neuron Number in Humans: Effect of Sex and Age. *Journal of Comparative Neurology*, 384(2):312–320, 1997.

[72] Derek Parfit. Personal Identity. *Philosophical Review*, 80(1), January 1971.

[73] Derek Parfit. *Reasons and Persons*. Clarendon Press, Oxford, 1984.

[74] Christopher Peacocke. *Sense and Content*. Clarendon Press, Oxford, 1983.

[75] Charles Sanders Peirce. *Collected Papers*, vol. 6 of *Scientific Metaphysics*. Harvard University Press, Cambridge, Mass., 1898, 1935. Charles Hartshorne and Paul Weiss, eds.

[76] Roger Penrose. *The Emporer's New Mind*. Oxford University Press, Oxford, 1989.

[77] Roger Penrose. *Shadows of the Mind*. Oxford University Press, Oxford, 1994.

[78] Steven Pinker and Alan Prince. On Language and Connectionism: Analysis of a Parallel Distributed Processing Model of Language Acquisition. *Cognition*, 28:73–193, 1988.

[79] Karl R. Popper and John C. Eccles. *The Self and Its Brain*. Springer-Verlag, Berlin, 1985.

[80] Hilary Putnam. Brains and Behavior, in Ned Block, ed., *Readings in Philosophy of Psychology*, vol. 1. Harvard University Press, Cambridge, Mass., 1980.

[81] Anthony Quinton. The Soul, in John Perry, ed., *Personal Identity*. University of California Press, Berkeley, 1975.

[82] Thomas Reid. Of Identity, in John Perry, ed., *Personal Identity*. University of California Press, Berkeley, 1975.

[83] Thomas Reid. Of Mr. Locke's Account of Our Personal Identity, in John Perry, ed., *Personal Identity*. University of California Press, Berkeley, 1975.

[84] Georges Rey. What's Really Going On in Searle's Chinese Room. *Philosophical Studies*, 50:169–185, 1986.

[85] David Rosenthal. Two Concepts of Consciousness. *Philosophical Studies*, 94(3):329–359, 1986.

[86] David Rosenthal. A Theory of Consciousness, in Ned Block, Owen Flanagan, and Güven Güzeldere, eds., *The Nature of Consciousness*, pp. 729–753. M.I.T. Press, Cambridge, Mass., 1997.

[87] David E. Rumelhart and James L. McClelland. P.D.P. Models and General Issues in Cognitive Science, in *Parallel Distributed Processing*, vol. 1, pp. 216–271. M.I.T. Press, Cambridge, Mass., 1986.

[88] Gilbert Ryle. *The Concept of Mind*. Penguin Books, Harmondsworth, England, 1949.

[89] Arthur L. Samuel. Some Studies in Machine Learning Using the Game of Checkers, in Edward A. Feigenbaum and Julian Feldman, eds., *Computers and Thought*. McGraw-Hill, New York, 1963.

[90] Jonathan Schaeffer. *One Jump Ahead: Challenging Human Supremacy in Checkers*. Springer-Verlag, New York, 1997.

[91] Jonathan Schaeffer, Joseph Culberson, Norman Treloar, Brent Knight, Paul Lu, and Duane Szafron. A World Championship Caliber Checkers Program. *Artificial Intelligence*, 53(2–3):273–290, 1992.

[92] Roger C. Schank and Robert P. Abelson. *Scripts, Plans, Goals, and Understanding.* Lawrence Erlbaum Press, Hillsdale, New Jersey, 1977.

[93] John R. Searle. Minds, Brains, and Programs. *Behavioral and Brain Sciences*, 3(3), 1980.

[94] John R. Searle. *Intentionality: An Essay in the Philosophy of Mind.* Cambridge University Press, Cambridge, 1983.

[95] John R. Searle. *The Rediscovery of the Mind.* M.I.T. Press, Cambridge, Mass., 1992.

[96] John R. Searle. *The Construction of Social Reality.* Free Press, New York, 1995.

[97] John R. Searle. *The Mystery of Consciousness.* New York Review of Books, New York, 1997.

[98] John R. Searle. *Rationality in Action.* M.I.T. Press, Cambridge, Mass., 2001.

[99] Charles P. Siewert. *The Significance of Consciousness.* Princeton University Press, Princeton, New Jersey, 1998.

[100] Peter Singer. *Animal Liberation.* New York Review of Books, New York, 2nd edition, 1990.

[101] J. J. C. Smart. Sensations and Brain Processes. *Philosophical Review*, 68:141–56, April 1959.

[102] Kim Sterelny. *The Representational Theory of the Mind.* Blackwell, Oxford, 1990.

[103] George Frederick Stout. The Nature of Universals and Propositions, in Charles Landesman, ed., *The Problem of Universals*, pp. 154–166. Basic Books, New York, 1921, 1971.

[104] Leopold Stubenberg. *Consciousness and Qualia*. John Benjamins, Amsterdam, 1998.

[105] Norman Stuart Sutherland. *The International Dictionary of Psychology*. Continuum, New York, 1989.

[106] Astro Teller. *Exegesis*. Vintage Books, New York, 1997.

[107] James T. Townsend. Don't be Fazed by PHASER: Beginning Exploration of a Cyclical Motivational System. *Behavior Research Methods, Instruments and Computers*, 24:219–227, 1992.

[108] Alan Turing. Computing Machinery and Intelligence. *Mind*, 59:433–460, 1950.

[109] Michael Tye. A Representational Theory of Pains and Their Phenomenal Character, in James E. Tomberlin, ed., *Philosophical Perspectives*, vol. 9. Ridgeview, Atascadero, Calif., 1995.

[110] Timothy van Gelder. Dynamics and Cognition, in John Haugeland, ed., *Mind Design II*, pp. 421–450. M.I.T. Press, Cambridge, Mass., 1997.

[111] Robert van Gulick. Understanding the Phenomenal Mind: Are We All Just Armadillos?, in Martin Davies and Glynn Humphries, eds., *Consciousness: A Mind and Language Reader*. Basil Blackwell, Oxford, 1992.

[112] Kendall L. Walton. *Mimesis as Make-Believe: On the Foundations of the Representational Arts*. Harvard University Press, Cambridge, Mass., 1990.

[113] Lawrence Weiskrantz. *Blindsight: A Case Study and Implications*. Oxford University Press, Oxford, 1986.

[114] Lawrence Weiskrantz et al. Visual Capacity in the Hemianopic Field Following a Restricted Occipital Ablation. *Brain*, 97:709–728, 1974.

[115] Joseph Weizenbaum. ELIZA: A Computer Program for the Study of Natural Language Communication between Man and Machine. *Communications of the* ACM, 8:474–480, 1967.

[116] Alan R. White. *Attention*. Basil Blackwell, Oxford, 1964.

[117] Bernard Williams. The Self and the Future. *Philosophical Review*, 79(2), April 1970.

[118] Donald C. Williams. The Elements of Being. *Review of Metaphysics*, 7:3–18, 1953.