# How the Brain Works:
# Explaining Consciousness

---

A Thesis
Presented to
The Division of Philosophy, Education, Religion and Psychology
Reed College

---

In Partial Fulfillment
of the Requirements for the Degree
Bachelor of Arts

---

Ben Salzberg

May 1994

Approved for the Division
(Psychology)

---

(Enriqueta Canseco-Gonzalez)

## Acknowledgements

Thanks go to Wendy Breyer for thesing with me, to Enriqueta Canseco Gonzalez for neuropsychological criticism, to Mark Hinchliff for reading my first chapter and for Philosophy of Mind, to Albyn Jones for help with statistics, to Zeke Koch for TrueBasic programming and positive reinforcement, to Aaron Mackey for statistics help far beyond the call of duty and then some, to George Mandler for the idea for my experiment, to Melinda Macpherson for encouragement and the use of her computer, to Nelson Minar for taking time off his own thesis to help me with mine, to Nick Rayner for reading the whole damn thing, to Dan Reisberg for inspiring me to think thoughtfully about thinking, to Alec Rogers for more TrueBasic programming and positive reinforcement, and to my parents for raising me curious. Special thanks go to my Grandparents Ted & Hilda Salzberg and to Robert P. Lamons for paying for the whole thing.

# Preface

> "Why,..." ask the people in Artificial intelligence, "do you waste your time conferring with those neuroscientists? They wave their hands about 'information processing' and worry about where it happens, and which neurotransmitters are involved, and all those boring facts, but they haven't a clue about the computational requirements of higher cognitive functions." "Why," ask the neuroscientists, "do you waste your time on the fantasies of Artificial Intelligence? They just invent whatever machinery they want, and say unpardonably ignorant things about the brain." The cognitive psychologists, meanwhile, are accused of concocting models with neither biological plausibility nor proven computational powers; the anthropologists wouldn't know a model if they saw one, and the philosophers, as we all know, just take in each other's laundry, warning about confusions they themselves have created, in an arena bereft of both data and empirically testable theories. With so many idiots working on the problem, no wonder consciousness is still a mystery. [Dennett, 1991]

What is a philosophical discussion doing in a psychology thesis? Is the history of the philosophy of mind at all useful to psychologists? I've started out with the philosophy of mind for several reasons, perhaps primarily because that's where I first started thinking about the global theories of brain activity, and how they fit

in (or did not fit in) with the philosophies I encountered. Initially, of course, I read of the mind/body problem as framed by Descartes, then John Searle's Chinese Room thought experiments , then a few epiphenomenal accounts, until I finally ended up with ideas based on a mixture of cognitive psychology and neuroscience, solidly grounded in physical reality. This progression is not obvious, though it may seem so from an enlightened perspective, and I believe that at the current state of psychology we can use all the help we can get. Philosophy can help to illuminate some of the mistakes we've been making over the years, and can help to point the way to the problems we need to solve. Through the process of identifying, isolating, and comparing philosophical standpoints, we can both give ourselves a strong ontological base and see where the unsolved mysteries really lie. Through the lens of philosophy we can examine psychology, see what assumptions have been made and how these assumptions change the interpretation of results. Although philosophers have a bad name in the sciences (they sit around in their armchairs and make ridiculous arguments without actually testing anything) the methods they use can penetrate some of the murkier issues of interpretation of results and data, an area in science which has too often been subjective. An enlightened philosopher, having studied both past attempts to explain the mind and current research, may be able to point out where scientists have been unknowingly holding on to old ideas, reinterpret unexplained data and highlight gaps in theoretical models. By examining several different ways of thinking about the mind we will be able to move beyond our old limitations and actually get to work on the study of consciousness. The purpose of this thesis is to demystify consciousness, to make it conceivable as a physical phenomenon. In order to do this I will show that many of the approaches and conceptualizations of consciousness thus far are based on either faulty philosophical or scientific assumptions. Whatever is left will at least point toward a workable, satisfying and understandable explanation. This endeavor need not be put off until

we have a completed neuroscience, or physics, or cosmology. I hope I can make my arguments and explanations clear enough that creating a model of how consciousness works will not look hopeless anymore, and show that empirical research can reveal some of the mechanisms of consciousness today, rather than in some imagined future.

# Contents

# List of Figures

# List of Tables

**Abstract**

This is a thesis in two parts. The first and largest part is an attempt to explain consciousness. This theoretical part includes a chapter on the philosophy of mind; a chapter on the findings and theories of consciousness from neuroscience, cognitive neuropsychology, cognitive psychology and information science; and a chapter reviewing Daniel Dennett's attempt at synthesizing all of these fields after which my own conclusions are stated. Consciousness is determined to be explainable wholly in terms of physical phenomena, many of which have been described by psychology.

The second part describes an experiment meant to test whether pre-experiential or post-experiential information has a greater effect on judgement, in order to test whether Dennett's Multiple Drafts theory or one of two contrasting theories (Stalinesque or Orwellian) were true. In this experiment subliminal stimuli (color words) were presented tachistoscopically along with supraliminal stimuli (grey shapes) to determine whether the subliminal stimuli affected subjects' judgement of the "color" of the supraliminal stimuli. No significant results were found. Suggestions for future research on this topic and on language's relation to consciousness are presented.

# Chapter 1

# Introduction

What is consciousness? What is it for? Is it a real phenomenon, something describable, or is it completely ineffable, only fit for study by pseudo-scientists and philosophers? Do we know enough now to be able to address these questions, or must we wait, perhaps for a completed physics[1] or a completed neuroscience?[2] Or can we even comprehend the mechanisms of the brain at all?[3] Since to fully answer all of these questions would be too mammoth a task for one year, I have restricted myself in some (perhaps controversial) ways. There will be no detailed metaphysical theses, and few physical details, as much work must be done before such an explanation of consciousness can come about. Nevertheless, I am confident that the ideas presented herein represent some progress in how we can think about consciousness, and will point to some ways of rooting out the deeper problems in our conceptions. By examining how philosophy and three of the branches of psychology

[1] As Penrose [Penrose, 1989] believes: Consciousness is too great a mystery right now, because it may depend heavily on quantum effects, inexplicable until we have (maybe) a Correct Quantum Gravity theory.

[2] Philosophical eliminativists (like [Churchland, 1981]) believe that once we have a better understanding of the low-level brain processes that lead to phenomena like emotions and consciousness, these questions will go away, these terms will go away, replaced by more specific physical descriptions (like "I'm feeling a little over-dopaminergic today"?)

[3] Philosopher Colin Mcginn believes that we may be "cognitively limited" with respect to consciousness, just as dogs are cognitively limited with respect to chemistry [McGinn, 1991]

have approached consciousness, I hope to point out many of the limitations that must constrain any explanation of consciousness.

Following Dennett's [Dennett, 1991] and Crick and Koch's [Crick and Koch, 1990] lead, here are some of the assumptions that I have used to limit my investigations:

1. There is something to explain: Consciousness, at least as a perceived phenomenon, exists, in the real world, which also exists. Metaphysical points seem to arise when the subject is consciousness, but there is no room for such a discussion here.

2. No magic allowed. This includes any supernatural explanations of consciousness, as well as speculative physics. No backwards-in-time causation, no obscure quantum effects, and no mysterious undiscovered forces. The point is to see if we can understand consciousness with the knowledge we have, rather than hoping for some magical intercession to save us from our conceptual difficulties.

3. All of the subtleties must be preserved. Certainly, subjective experience is a rich and complicated thing, and any explanation which leaves out parts of it is somehow lacking. Included here are qualia, the apparent special qualities ("raw feels") of experience and why it is like something to be a subjective experiencer. Qualia are what we have and rocks and trees don't, they are what it feels like to stub your toe, or see a sunset, they are the kinds of experiences that things that it is "like something to be like" have.[4]

4. Any good model of consciousness must use what we already know about the mind/brain. It should use as many of the results of current and past research on the brain as possible as a base from which to construct a model, rather than starting with a model and from there constructing how the brain must work. There will be very little questioning of the scientific method, as I believe that we haven't yet

[4]There is a further discussion of this idea in the first chapter, under the heading "property dualism."

exhausted the means at our disposal in the study of consciousness. Using what we already know about the mind/brain entails, I think, believing that the brain (the physical organ) produces consciousness.

## 1.1 Philosophy of Mind

I will examine several of the major trends in the philosophy of mind in the first chapter, and attempt to point out for each its most attractive features as well as its most difficult problems.[5] This examination will show that the views of Descartes, Nagel, Jackson, McGinn, and Penrose (among others) leave no room for scientific investigation into the nature of consciousness, and that according to them a scientific explanation of consciousness is impossible in principle. I will discuss substance dualism (Descartes, Penrose[6]), property dualism (Nagel, Jackson, Searle, McGinn) epiphenomenalism (Velmans, Nagel, Jackson), and physicalism (P.S. & P.M. Churchland, Dennett). All of the above philosophies save physicalism leave a residual mystery, unexplained and unexplainable. I will also discuss three more current philosophical stances which do allow room for scientific investigation: Functionalism, type physicalism, and token physicalism. All of these positions can accommodate the physicalist notion that the brain is the mind in some way, although their ways of explaining what the mind is differ in important ways.

[5] Philosophers of mind may be disappointed that I have not gone into sufficient depth in chapter 1 (and even that I have omitted some important positions); let it be said here that I am not attempting a detailed philosophical thesis, I am merely trying to get a grip on some of the philosophical problems involved.

[6] Penrose is a physicist, but believes that some fundamental 'other' mechanism is needed to explain consciousness.

## 1.2 Psychology of Mind

The second chapter will give a brief overview of the accomplishments of the empirical research into the structure and functions of the mind/brain. In each of the three sections (on neuroscience, neuropsychology and cognitive psychology) I will discuss the major findings, methods, and theories of consciousness. How have psychologists approached questions of consciousness? One way of dealing with these questions is to deny that there is a problem (a mind or consciousness) at all, as the "barefoot behaviorists"[7] did earlier in the history of psychology. If we deny that mental phenomena have any significance, or even deny their existence, then the problem disappears. Stimulus-response learning is an important part of explaining behavior, but is by no means the whole story.

Neuroscientists on the other hand are concerned purely with low-level mechanisms and traditional scientific techniques. They have taught us much about the mechanisms of the brain, even some of the functional architecture of the brain, and yet they also tend to shy away from questions like "What is consciousness?" This level of explanation can tell us about which brain areas are strongly connected to which other brain areas, about what kinds of neurotransmitters exist, about different kinds of synaptic connections, but not enough about larger mechanisms. At this level of explanation, we cannot ask "What is thought?" and receive an understandable answer. A list of chemical interactions or of which neurons are firing is not understandable—unless tied to a larger scale picture, it is just a laundry list of facts.

At a broader, less fine-grained level of explanation is cognitive psychology. Cognitive psychology, like behaviorism, examines the inputs and outputs of the mind,

[7]Extreme behaviorism means denying the possibility of explaining mental processes: the mind is an unassailable "black box," and all human actions can ultimately be explained by stimulus-response behaviors.

but then uses this information to make hypotheses about the structure of the mind, rather than saying that the stimuli and responses are all that cognition is. While this field can tackle the kinds of questions we want answered, cognitive psychologists have taken another way around, by using a divide - and - conquer strategy for explaining cognitive mechanisms and carefully avoiding use of the word "consciousness." Instead, there are "auditory filters," "rehearsal loops," "central executives"—all of which perform some of the tasks the human brain accomplishes. While these are useful ideas, and have taken thought about the brain to new levels, they still do not explain the residual mystery: How does this collection of mechanisms become a thinking being, a consciousness? Why is it like something to be me?

Evidence from cognitive psychology (which usually takes a functionalist philosophical stance), neuroscience (which usually takes a type or token physicalist stance), neuropsychology (also type or token physicalist) and computer imaging (used by all three fields of psychology) has taught us much about the brain—that the brain must use some kind of parallel distributed processing to accomplish what it does, that certain sections of the brain appear to have highly specialized functions, that some sensory processing areas have pandemonium-type architecture, that some areas of the brain may be informationally encapsulated, functioning as separate processors, not to mention all of the psychopharmacological and neurotransmission findings[8]—an enormous quantity of information, from thousands of studies. Using the philosophical base developed in chapter one, we may reinterpret some of the experiments which bear most directly upon consciousness. Do we have enough information to understand how the brain works? Are we wise enough to understand what the experiments we have done imply? Can we build a theory of consciousness on this foundation, or are we missing crucial pieces? What kind of explanation will satisfy our wish to understand consciousness? These are the questions answered by

[8]These topics are discussed in detail in chapter 2.

the third chapter, in which many of the old models are seen to lead to the Cartesian Theater (an understanding of the mind which includes a homunculus, leading to infinite regression [Dennett, 1991]) or to unsatisfactory types of explanations.

The problems with any theory that includes a Cartesian Theater will be explained, and an alternative model will be put forth: Dennett's Multiple Drafts model, wherein there is no seat of consciousness within the brain, no place where it all comes together, but rather a mass of individually limited systems all acting and interacting at once. Consciousness does not come from an immaterial world as the dualists believe, or emerge without purpose or causal power as the epiphenomenologists believe, or sit in the Cartesian theater watching the show and making clever decisions, as "homuncularists"[9] believe. By the time many of your brain's clever decisions become conscious, they have already been made. This does not mean that you didn't cause them—it is merely further evidence that introspection does not bring true insight to the workings of the mind, that you don't really know intuitively how your brain works. Dennett's book is huge and multipartite, so some of its richness will be lost in this summary. However, the most important points (and their respective justifications) will be presented and criticized, along with my own conclusions.

Finally, the last chapter will present an experiment meant to test a part of Dennett's theory, with regards to pre-conscious memory editing and alteration of perception. There are two hypotheses about how these kinds of perceptual/memory revisions occur, Stalinesque and Orwellian.[10] The Stalinesque theory proposes that conscious memory of events is altered by pre-experiential information, so that the eventual experience is a kind of show trial: the outcome is determined by what has come before. The other theory, Orwellian revision, says that the perception

[9]Meaning "those whose theories contain an unexplained homunculus."
[10]These terms are from [Dennett, 1991]

is changed after the experience—history is rewritten, so to speak, by an Orwellian editor before it becomes conscious. Dennett believes that no distinction can be made between these two theories, that instead "Multiple Drafts" of the experience are simultaneously active. If there is a distinction, then one of the theories seems to imply pre-conscious perceptual editing, and the other post-conscious, and there is therefore some particular moment of consciousness, perhaps in the Cartesian theater [Dennett, 1991]. The experiment presented in the last chapter is an attempt to determine whether pre-perceptual or post-perceptual influences have more effects on perception, and to look for evidence of either kind of editor.

# Chapter 2

# Philosophy of Mind

In this chapter I will present the mind/body problem as philosophers have approached it. I will examine only six of the most important philosophical positions, and each of them fairly briefly, but these positions should demonstrate the ideas philosophy has to offer, the solutions it has proposed, and the questions that still remain.

Keith Campbell, in Body & Mind, [Campbell, 1970] gives a set of four statements that form an "inconsistent tetrad," in which any three can be logically accepted but in so doing make the fourth unacceptable. In his words, they "express in a nutshell the dilemma which confronts us." The tetrad is:

1. The human body is a material thing.

2. The human mind is a spiritual thing.

3. Mind and body interact.

4. Spirit and matter do not interact. (pg. 14)

Any three of these propositions may be true, but together they will be inconsistent

with the fourth[1]. To solve the mind/body problem is to prove that one of the four propositions is false. In this thesis I will take the position entailed by (1), (3), and (4), which rule out (2) and entail instead:

5. The human mind is not a spiritual thing.

Once we have decided that the mind is a physical thing, or at least results from physical processes in our brains, we will be able to study it systematically, and eventually we will come to understand consciousness. In support of (5) we will examine several philosophical positions which in some way contradict it, namely substance dualism, property dualism, and epiphenomenalism; and some which support it: functionalism, and type and token physicalisms. In each of these sections only a few of the most important arguments will be examined, enough to impart the main ideas and objections without going on too long. Of course there will be omissions that will seem glaring to some, but this chapter is not meant to be an exhaustive review of the philosophy of mind. Rather, the intent is to show a few of the ways the problem of consciousness has been attacked in the past so we can learn from their successes and failures and move on.

## 2.1 Substance Dualism

One of the first important figures in the study of consciousness is René Descartes, at least because it is with his ideas that many of us first thought about thought. René Descartes was a supreme mechanist, about most things. He understood the complex machinery behind the animated water sculptures popular at the time. He lived during an age of mechanism, when God was seen as the Great Watchmaker. He came up with a mechanism for reflex action that was remarkably good, though

[1]Campbell gives an example (pg. 15): "(1), (2), and (4) entail another statement, (5) Mind and body do not interact, and (3) and (5) together are a flat contradiction."

wrong. And yet, for Descartes as for many people today, the mind was something different. Mysterious, complex, unexplainable, immaterial. Clearly, it was made of a different kind of stuff, a nonphysical stuff. How could the feeling of a cool breeze (the beauty of a Mozart sonata, the fear of a barking dog, etc.)[2] be reduced to mere physical mechanisms? They are of different *kinds.*

This insight was the result of a long period of introspection, wherein he wondered whether all that he believed could be false. This questioning technique led to some serious consequences, all based on what could be doubted. Have we all not misheard someone else's speech? Then perhaps the information passed on by our ears can be doubted. Couldn't a square tower appear round from a distance? Then perhaps we should doubt our eyes. Likewise with all the other senses: none of them is supremely reliable. Of what, then, could he be sure? If all his senses could mislead him, if some evil demon was feeding him consistently false information[3], then could he doubt everything? No, for the very act of doubting proved to him that he existed, and that at least was certain. *Cogito ergo sum.* Since everything that was physical could be doubted (because all of his senses could be fooled, or he could be dreaming) this thing that he was, this thinker, had to be a separate *kind* of thing, a nonphysical kind of thing. There must be two kinds of stuff: immaterial mind stuff and material stuff.

Descartes needed a way for this mind stuff to interact with the physical brain stuff, because it certainly seems like our thoughts and decisions influence our physical bodies. His candidate for the point of interaction was the pineal gland, an organ

[2]These qualities, the sort of unique experiential things humans enjoy, are called *qualia* by philosophers.

[3]This circumstance could be imagined in a more sophisticated way. We could suppose a truly realistic virtual reality setup, where the evil computer programmer composes an artificial reality. While this is possible in principle, it is not so in fact: modeling even a very simple world, with just a few details and very limited interactability requires truly massive computing power, and wouldn't really fool anyone. Adding realism requires geometric increases in computing power, so that simulation becomes impossible in fact. (For more see [Dennett, 1991] pp. 6-7 and note.)

in the center of the brain. Impulses from the immaterial mind causally interacted with the pineal gland, thus causing the nerve fibers to activate the body. This philosophical standpoint is a species of substance dualism: There are two distinct *kinds* of substance, which interact in some way. This can be an attractive view because it keeps our sense of specialness, of being a different kind of thing than the physical world. We can have an eschatology, immortality, transcendence. It puts the mysteries of the human brain into place as inherently mysterious: while we can understand how physical systems can have reactions, stimulus - response interaction with the world (as Descartes believed was the case for animals ([Carlson, 1991] p. 3)), it is difficult to imagine how any such physical system could have will, intentions, goals, let alone things like the appreciation of a beautiful sonata or the comfort of a warm bed (*qualia*).

Many of us have not yet abandoned this sixteenth-century viewpoint, tenaciously hanging on to enigmatic theories of mind, perhaps unknowingly protecting something we wish to keep sacred. But abandon it we must, for its problems are numerous and intrinsically unsolvable. How does this mind stuff interact with the brain? Where is it? Is it a detectable substance? If its existence cannot be proven in any satisfying way, then why should we believe in it? I hope these questions are enough to cause one to doubt this model of the mind[4]. Wouldn't we be more satisfied if we could explain the workings of consciousness without resorting to a magical, unobservable kind of stuff? Accepting substance dualism shuts off empirical research that seeks to understand consciousness; it says that there is an entirely separate kind of

[4]Formal philosophical objections to Descartes' modal and certainty arguments exist. The modal argument, briefly stated, says (1) $\diamond$ (I exist without my body). (2) If (I) is me, I $\neq$ my body. (3) Therefore, I $\neq$ my body. The first premise is troublesome, as Descartes goes from conceivability to possibility (it is conceivable that I exist w/o my body $\rightarrow$ it is possible that I exist w/o my body) using the existence of God to back him up. (clear and distinct perception, and therefore conceivability, come from God and imply possibility. (See Descartes Meditations on First Philosophy for more.

stuff that we can't investigate. As Dennett says, *"accepting dualism is giving up."* ([Dennett, 1991] p. 37) If our curiosity demands that we understand consciousness more thoroughly, we have to abandon substance dualism as a dead-end route. While it is true that we don't yet have any such satisfying materialist explanation of consciousness, we shouldn't give up hope. There is still much to be discovered—we have not yet exhausted the means at our disposal. Scientists cannot accept magic, but must examine and explain it.

## 2.2 Property Dualism

Another kind of dualism which exerts perhaps the most subtle and tenacious hold on theories of mind is *property dualism*. Property dualism says that mental events are distinct from other kinds of events because they have a separate kind of property, something special, which emerges from the action of the brain ([Levine, 1993]). This standpoint often leads people [Searle, 1984] to argue that consciousness, as experienced by humans, is a unique kind of thing which leads to the conclusion that even a complete understanding of the structure and function of the brain cannot get at the special property of consciousness[5]. Property dualists often say they are strict materialists [Searle, 1984]—no one wants to admit to substance dualism, yet they wish to reserve some specialness for consciousness. While insisting that they are not dualists, they are committed to a kind of dualism by their tenacious desire for some mysterious property or their inability to see that this is where their arguments lead: We cannot be materialists except for the special property of consciousness.

Arguments for the special property of consciousness are often intuitively convincing, but actually turn out to be misleading. Nagel's "What Is It Like to Be a Bat?"

[5]This philosophical stance has much in common with epiphenomenalism, because consciousness becomes something separate from the physical workings of the brain, except that epiphenomenalists don't have to be monists—they can believe in a dualistic mind, and not have to worry about mind-matter interaction.

[Nagel, 1974] emphasizes the intrinsic unsharable quality of subjective experience. This idea is quite gripping, because it gets at one of the most mysterious qualities of consciousness: there are some things that it *is like something to be,* as opposed to things it is not like something to be, and we can't know what it is like to be someone else. We know, from introspection, that we exist (*cogito ergo sum*) and that it is like something to be us. We can also be fairly sure that is not like anything to be a rock, or even a tree[6]. What is this magic property of consciousness? According to the property dualists, a physicalist explanation of the brain, even if complete, would "leave something out."[7] Specifically, the property of consciousness cannot be reduced to a physical state, or even physical processes.

The next example illustrates a similar idea, that there is something about subjective experience unexplainable by science. It is Frank Jackson's "knowledge argument," against materialism, and is the story of Mary, "a brilliant scientist." [Jackson, 1982] Mary is kept for her whole life in a black and white world, so she has no experience of color whatsoever. She learns all the physical information there is to know about vision, everything that can be known about seeing a red tomato, or a yellow banana, or the sky at sunset. After she has acquired all this knowledge, she is released and experiences color. What happens? Does she learn anything? It seems obvious that she will, and yet she has all the physical information there can be. Therefore, a physical explanation is not all there is, so we cannot reduce experience to physical properties. There must be something else, and so we are led

[6]This problem brings the older theories of life to mind: we have no need, now, to resort to the *élan vital* to explain what makes living stuff alive, but the question "What is life?" remains difficult. Is there something intrinsic about being alive that science can't explain? Is there something intrinsic about subjectivity that science can't explain? Some things we are sure about, like rocks vs. trees, but some are more vexing, like computer viruses or artificial life programs. The question of animal consciousness will unfortunately not fit here, but is still a problem: Are gorillas conscious? Cats? Ants?

[7]Levine, in [Davies and Humphreys, 1993]: "On Leaving Out What it's Like." (from the context, it sounds like I'm saying Levine takes this position. He doesn't, but I liked his phrase.)

to another kind of dualism, property dualism.

There are several different ways to reply to Nagel's and Jackson's arguments. The simplest is to dismiss them outright: since they both lead to dualism, and accepting dualism means accepting that we cannot have a viable scientific explanation of consciousness, we may ignore them, because we are not yet ready to give up. [8] One reply to Nagel's question is this thesis, and others like it: if we understand why and how the thoughts that come into our heads come into our heads, and why we can't introspect the mechanisms underlying this process, then we will have some understanding of our unique subjective viewpoint—why it is *like* something to be a human.

The replies to Jackson's argument either attack the concluding intuition (Levine, in [Davies and Humphreys, 1993]) or the premises of the argument itself ([Dennett, 1991] and P. M. Churchland in [Dennett, 1991]). Levine (following Horgan 1984a) shows that the fact that Mary learns something new does not mean that materialism is bankrupt. Knowing (even in the most precise physical terms) something and experiencing something are epistomologically distinct, but both can be described physically. Even knowing all there is physically about how the process works would not cause that process to happen (i.e. knowing about the breakdown of various photopigments in the retina is different from experiencing it[9]). Churchland and Dennett attack the intuitive feeling that one can know everything physical there is to know about seeing red: it would be an enormous pile of information, and if she could know introspectively (or find out through the same methods with which that knowledge was presumably garnered) what that experience was, then she would not

[8] This is convincing enough for me, but facile, because any successful explanation of consciousness should not only be able to say why these arguments are wrong but also why they are so seductive.

[9] A peasant in 16th century England might know about lightning, and experience it, but not know anything at all about electricity. Nonetheless, lightning *is* electricity. (U.T. Place, "Is consciousness a brain process?" in [Lycan, 1990])

learn anything new. The fact that we cannot introspect the workings of our own brains is well established—but if we could (if Mary could) and we knew what Mary is supposed to know, then we would learn nothing new.[10]

## 2.3 Epiphenomenalism

Another philosophical stance, one that has much in common with property dualism, is *epiphenomenalism*. Epiphenomenalism says that the mind, consciousness, is epiphenomenal with respect to the processes and actions of the brain. Dennett [Dennett, 1991] makes a distinction between the philosophers' more strong definition of epiphenomenalism and the scientists' more limited definition. One problem, he says, is that these two different definitions apply to the same word, which is used by both parties as if they mean the same thing.[11] When psychologists describe something as being epiphenomenal, they mean that that thing is nonfunctional. When philosophers call something epiphenomenal they mean that it is an effect, but "has no effects in the physical world whatsoever." (pg. 402) So, for a scientist, the noise and heat created by an engine as it drives a car are epiphenomenal, incidental to the function of the car. But the epiphenomena are still effects on the physical world: they can be sensed (or recorded by instruments). These are not the philosophers' epiphenomena. For them, the phenomena must have no effects whatever on the physical world. ([Dennett, 1991] pg. 402) This is a much stronger statement: if consciousness is an epiphenomenon, we cannot measure it, divine its contents, probe its structure—for it has no physical correlates. The attraction of

[10]Dennett provides an amusing continuation of the story of Mary ([Dennett, 1991] p399-400): At the end of Mary's incarceration, she is allowed to see a banana—which her captors have colored blue! She says "Hey! bananas aren't blue!" How did she do it? She knew what it would be like to see yellow, because she knew *everything* about seeing, and since her expectation did not meet up with experience, she saw through the trick.

[11]This kind of low-level miscommunication can be found wherever philosophy and psychology interact, and unfortunately even within these fields.

philosophical epiphenomenalism is its unprovability—its safety from scientific proof or disproof. It can preserve beyond criticism the belief that there is something inherently special about the human mind, about consciousness. It can be an argument that artificial devices cannot have consciousness: since they don't have this special epiphenomenal mind, they are not the same as us. The problem with this stance is that it is unprovable. Here is a case where we have strong grounds for being verificationists, because epiphenomena lead to cases where everyone should want to be verificationists. ([Dennett, 1991] p. 403)[12] Dennett provides an example here:

> Consider, for instance, the hypothesis that there are fourteen epistemological gremlins in each cylinder of an . . . engine. These gremlins have . . . no physical properties; they do not make the engine run smoother or rougher, faster or slower. There is *and could be* no empirical evidence of their presence, and no empirical way in principle of distinguishing this hypothesis from its rivals: there are twelve or thirteen or fifteen . . . gremlins. ([Dennett, 1991] p. 403-4)

This seems arbitrary, ridiculous. But why is it more ridiculous than the idea of an epiphenomenal consciousness? Why should we accept it, if it is presented in the same terms? Consider the hypothesis that mind is made of a different kind of stuff than brain, that it is essential to consciousness but non-physical and unobservable and doesn't interact with the body/brain physically. What do we lose by rejecting this kind of idea? Our tradition of accepting qualia as magical, separate, epiphenomenal? If the very idea of a non-physical mind is rendered completely impotent by those who wish to preserve it, we are led to wonder what it is that they are preserving.

> And what if some benighted people have been thinking for generations

[12] And on p. 461: ". . . if we are not urbane verificationists we will end up tolerating all sorts of nonsense: epiphenomenalism, zombies, indistinguishable inverted spectra, conscious teddy bears, self-conscious spiders."

> that gremlins made their cars go, and by now they have been pushed back by the march of science into the forlorn claim that the gremlins are there, all right, but are epiphenomenal? . . . These are not views that deserve to be discussed with a straight face. ([Dennett, 1991] p. 404)

What about the more serious hypothesis that consciousness is epiphenomenal in the scientific sense? If consciousness is real, and epiphenomenal, then it has no causal powers. It is merely an unconnected observer lurking in the brain, watching things happen but causing nothing. Evaluation of this view can only come after we understand more about what it means, how the physical architecture of the brain is set up, some of the parameters of its functions, and the arguments against any kind of passive or active observer watching the brain's machinations. But first let us turn to some philosophical positions that can support a scientific theory of consciousness.

There are three important philosophical positions which allow for scientific examination of the mind/brain: Type physicalism, token physicalism, and functionalism. While not mutually compatible or mutually entailing, each of these stances preserves many of the assumptions I would like to preserve, namely that consciousness is explainable, that the brain is what causes (or is) the mind, and that empirical science can help find an explanation of consciousness.

## 2.4 Type Physicalism

Type physicalism,[13] in one sentence, is simply the belief that every mental property is a physical property. Pain is a favorite example for philosophers, and the physical property is usually called the neural state of "firing c-fibers."[14] So for the type

[13]Much of the following discussion of physicalism (as well as some of the discussion of functionalisms) comes from lectures by M. Hinchliff, and are taken from his 1994 class Philosophy of Mind (Phil. 315, Reed college).

[14]This example often causes neuroscientists to howl. A more precise example would be: pain = activation of the spinothalamic systems leading to the ventral posteromedial and ventral pos-

physicalist, the property of being in pain=the property of being in the neural state characterized by firing c-fibers. Each mental property (i.e. the sensation of redness) is identical with some neural property or state (e.g. excitation of the visual cortex in such and such areas). There is substantial correlative evidence available for the type physicalist, and their claims seem quite empirically testable. In mammals, for example, both natural and artificial stimulation of c-fibers cause pain, and a mammal determined to be in pain exhibits firing c-fibers (determined by single cell recordings, perhaps).

There is a problem with this description of mental states, however, known as the problem of multiple realizations. If the property of pain just is the property of having firing c-fibers, what about organisms that don't have c-fibers? For example, squids could be examined, and found to behave as if they had the property of being in pain. Upon close examination of their nervous systems, no "c-fibers" are found. Does this mean that squids, contrary to the way they behave, do not ever have pains? The type physicalist surely does not want to say this. What about other kinds of creature? What if we were to come in contact someday with aliens who had strikingly different physiologies from ours (perhaps silicon-based[15]) and yet seemed to experience pains just as we did? Something is lacking in the type physicalist's explanation: what is in common between these disparate neural realizations of pain in virtue of which they are pains? The type physicalist response to this objection has been to say that the different neural states form a disjunctive set, so that pain, for example, is c-fibers in humans, q-fibers in squids, ... z-fibers in aliens. This is unsatisfying because it has not answered the objection properly: Yes, maybe it is those properties in those creatures, but what is it that they all have in common?

terolateral thalamic nuclei (sharp pain) or the parafascicular and intralaminar nuclei (deep or throbbing pain). ([Carlson, 1991] pp. 221-226) Since this is such a mouthful I'll stick to c-fibers, with the reader's understanding that it is out of convenience rather than ignorance or perversity.

[15] *Pace*, biologists and chemists. It's just an example.

What makes them all pain?

## 2.5 Functionalisms

The objection to type physicalism just raised often comes from functionalists.[16] Functionalists believe that what mental states have in common is the functional role they play. So pain is the property of having some property which has the functional role of pain. There are a variety of different kinds of functionalism, as this basic definition of functional roles allows for some leeway—one could even be a dualist and a functionalist at the same time, because that property which has the functional role of pain could be anything, even a soul-state. This metaphysical flexibility rescues functionalism from the problem of multiple realizations at the expense of any kind of specificity about the facts of the matter. Functionalism cannot be compatible with type physicalism, but it can be with token physicalism, discussed below.

There are at least two important kinds of functionalism, each of which insists that mental states are functional states but to varying degrees of specificity. Machine functionalism says that mental states are like the states of a turing machine, and that mental activity is governed by the kinds of inputs, machine-table states, and outputs described by a realization of a turing machine. Philosophers have troubles with machine functionalism because it seems to allow things that don't seem to have mental states to have them; hence it is too liberal and allows things like computers (or Coke machines!) to have mental states. Psychofunctionalism is slightly different from functionalism and machine functionalism: it says that mental states are psychological states, those states characterized by empirical psychology. According to Block, this kind of functionalism is too narrow—it allows only those things which

[16] [Block, 1980] "Troubles with Functionalism" pp. 268-305.

realize states functionally equivalent to human psychological states to have mental terms, although it seems plausible that there could be creatures with a different psychology who still had mental states (pp. 291-292). Only if the psychology referred to by psychofunctionalism is a kind of "universal psychology" will it be a good explanation of mental states, but Block is dubious about the possibility of such a comprehensive psychology (p. 292).

## 2.6 Token Physicalism

Token physicalism, in a sentence, is the belief that every mental event is a physical event. A "token" is just some physical manifestation of something, as opposed to a "type" which is a property of something. Token physicalism can be used to ground functionalism to the real world: for example, the functional state of pain is "tokened" as firing c-fibers in humans, but could be tokened in other ways (as long as it played the same functional role). But the physical instantiation is not the definition of the mental state: pain must be described at some other level ([Lycan, 1990] p.7).

There are some criticisms of token physicalism, mostly because of its lack of bold claims. Jerry Fodor [Fodor, 1986][17] says that it is weaker than materialism, type physicalism, and reductionism. Token physicalism alone is a weak position because it doesn't make any claims about natural kinds or scientific laws. It merely states that mental events are physical events which cause things to happen. It is silent about whether psychology will come up with laws to describe those mental events, about whether the kinds of mental events we talk about even exist[18] in the categories we group them into, and about whether the entities described by psychology can be reduced to physics.

[17] From an article reprinted in [Boyd et al., 1991].

[18] Eliminativists (like the Churchlands) actually believe that things like beliefs and desires don't actually exist as real things: they are remnants from "folk psychology" and should be replaced by more accurate terms from a completed neuroscience [Churchland, 1981]

After examining some of the main positions in the philosophy of mind I will reveal my prejudices and conclusions. Because I am curious and skeptical but willing to accept the proofs available through science,[19] I feel I must reject any theory which puts any part of consciousness as a phenomenon out of the sphere of things science can investigate. Any position which says that there are some things about consciousness which are impossible in principle to discover I find to be a discouraging and groundless one. How can we be sure, if science still has a few more techniques to use, if technology improves enough to investigate the brain in new ways, if cognitive psychologists are still discovering lawful explanations of human cognitive phenomena, that consciousness is incomprehensible in principle? Therefore I reject substance and property dualisms and epiphenomenalism, because all of them insist that consciousness cannot be explained. In some sense I believe that functionalism must be true: any explanation of consciousness must not be based on the physical manifestations of particular mental states, as these can be quite different in different organisms and even in different people. However, these mental states are physically realized, so I believe in token physicalism. The conjunction of some form of functionalism and token physicalism will, I believe, result in a satisfying and complete explanation of consciousness. To be convinced of this we must look at some of the ways in which mental states are tokened and the functional explanations of these physical facts. Now we are in a position to begin examining some psychological accounts of brain functions and consciousness, including those from neuroscience, neuropsychology, and cognitive psychology. After examining the products of those fields I will present Daniel Dennett's [Dennett, 1991] effort to explain consciousness, present my own attempts, and in the last chapter I will present an experiment meant to test some of Dennett's claims.

[19] Even if they turn out to be contingent and *a posteriori*.

# Chapter 3

# Neuroscience, Neuropsychology, Cognitive Psychology

These three fields of psychology have all approached the problem of consciousness at different levels of explanation. The findings, methods, and theories of consciousness developed by each subfield of psychology are distinct but overlapping; consequently each section of this chapter will explore these parts separately while emphasizing the similarities and differences between them. Neuroscience and neuropsychology have studied the physical brain: neurons and their interactions, groups of neurons as brain structures, and the functions of these brain structures as they relate to the brain as a whole. The neurosciences investigate the physical mechanisms of the brain, like neurotransmitters, synapses, axons and dendrites, receptor sites, and the complicated interactions among them. Cognitive psychologists have studied the functions and parameters of brain processes. Within this domain are memory, computation, emotion, decision-making, and language. All of the functions described in cognitive psychology are understood to be caused by entities described by neuroscience and neuropsychology, but none of them are explicitly mentioned. For example, there is a brain substrate for memory, but cognitive psychology does not describe it. Rather,

it describes the things that memory does, its limitations and strengths.[1]

Each of these fields uses different methods (although the distinctions are blurry at the edges: each field uses techniques used by others, like the behavioral testing of lesioned rats in neuroscience or the psychology-wide applications of computer-aided imaging) and so provides different kinds of information. Each have had some successes in explaining consciousness as well as some failures. Converging evidence from all fields of psychology will give us an idea of what a theory of consciousness must account for and how it can be explained without resorting to any of the philosophical dead ends discussed in the first chapter, and understanding some of the problems in each field will help us avoid mistakes when developing a better theory of consciousness.

[1]These discoveries often lead to progress in the neurosciences by stimulating research: the brain must accomplish certain functions, says cognitive psychology. How? asks neuroscience, and it proceeds to answer this question by reference to specific physical mechanisms.

## 3.1 Neuroscience

In the following two main sections I will sketch the findings and methods of neuroscience. First we will explore the physical structure of the brain and its constituent parts, ending with neurons and neurotransmission. Some of the methods of neuroscience which produced this knowledge will be explored as well. Understanding the fundamental complexity of the brain will lead us part of the way toward understanding how it can produce consciousness; understanding how we have learned what we have about the brain will provide some techniques for investigating consciousness.

### 3.1.1 Neuroanatomy: From Brain Areas to Neurons to Neurotransmission

Even at first appearance, as the earliest curious people looked at it, the brain is a complicated organ. The outermost and largest (in humans) part of the brain, the forebrain, is multi-lobed and contains many smaller divisions. The mid- and hindbrain are also complex. This organ is no simple pump, like the heart or filter, like the kidneys: it is a complex behavioral control center, an analytic machine, a survival tool, etc. It is also the organ whose actions create consciousness, a phenomenon still mysterious to us after we have figured out most of the complex machinery of the rest of the body. In this section I will very briefly sketch some of the main structures of the brain (forebrain, midbrain and hindbrain) and their main functions, as far as is known,[2] as well as a brief sketch of neuron and synapse structures and functions. Figures 3.1 and 3.2 should aid in visualizing these structures.

#### Major Brain Structures

The forebrain is divided anatomically into two parts, the telencephalon and the diencephalon. The telencephalon is composed of the cerebral cortex, the limbic system,

[2]Most of the following sketch of the brain comes from [Carlson, 1991] pp. 82-97.

and the basal ganglia. The cerebral cortex is the two outer symmetrical halves of the brain, what most of us think of as the brain. It is divided into the frontal (forehead area), parietal (top of the head area), temporal (sides of the head) and occipital lobes (back of the head). Each of these lobes performs some unique functions. The frontal lobes are thought to perform thinking and planning functions, to be able to access sensations and memories; basically, to think. Within the frontal lobes are the motor association cortex, the primary motor cortex, and Broca's area (part of speech production). The parietal lobes contains the primary somatosensory cortex and sensory association cortex, and is believed to store perceptions and memories and integrate sensory information into larger pieces. The temporal lobes contain Wernicke's area (another part of the language system) as well as processing auditory information. Perceptions, memories, and sensory association are thought to happen here, too. The occipital lobes process visual information and deal with perception and memory. The two halves of the brain are linked by the corpus callosum, which consists of axons linking analogous areas of either half of the brain.

The limbic system lies within the cortex, and its main structures are the hippocampus and the amygdala. These structures are important in emotions, learning and motivation. The hippocampus is thought to play a part in transforming short term memories or impressions into permanent or long-term memories ([Carlson, 1991] pp. 88, 480-510). "The basal ganglia are involved in the control of movement." ([Carlson, 1991] p. 88) They are located near the center of the head.

The second part of the forebrain, the diencephalon, is composed of the thalamus and hypothalamus. The thalamus is a kind of receiving station, taking in sensory information and projecting it to various cortical areas. Carlson says "most neural input to the cerebral cortex is received from the thalamus..." (p. 89). The hypothalamus controls the autonomic (self-governing) nervous system and survival

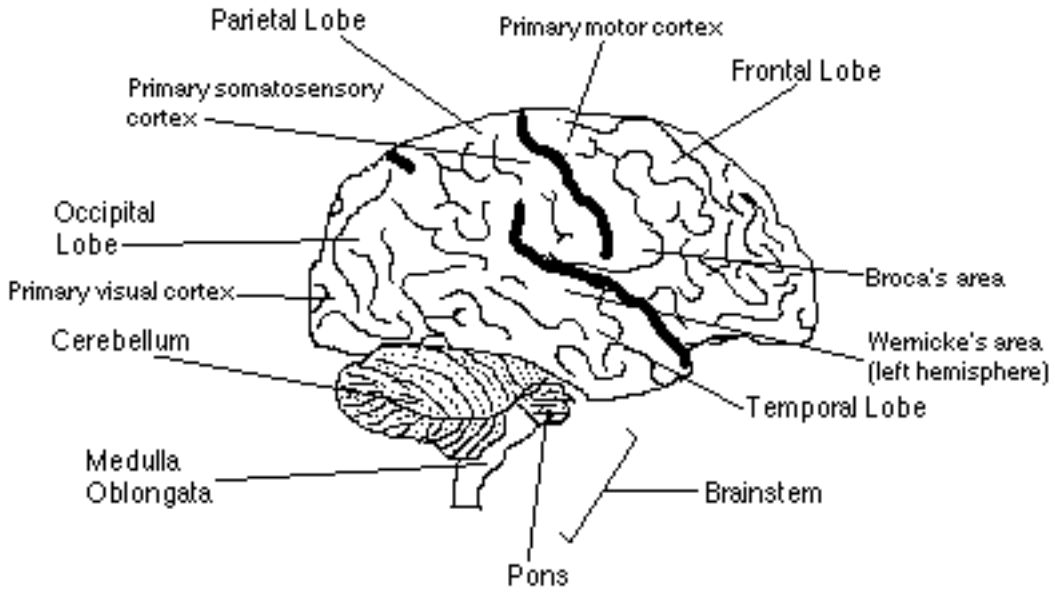


Figure 3.1: **Major Outer Brain Structures** (and some functional parts, Adapted from [Carlson, 1991] (passim) and [Nieuwenhuys et al., 1988] (passim).

functions, the "four F's: fighting, feeding, fleeing, and mating" (ibid.). Also within the hypothalamus is the pituitary gland, which is involved in hormonal control of many behaviors and sex differentiation and development.

The midbrain (mesencephalon) lies under the cerebrum and cerebellum, and along with the diencephalon and hindbrain forms the brain stem. It includes the inferior (lower) colliculi, part of the auditory system; the superior (higher) colliculi, part of the visual system; and the tegmentum. The midbrain is an older brain section evolutionarily, controls many important basic functions, like sleep, movement, species-typical behavior, and contains axons that connect motor systems of the brain to the spinal cord. The ventral tegmental area (which controls pleasure and reinforcement) is part of the midbrain.

The hindbrain contains the cerebellum ("little brain" ([Carlson, 1991] p. 94)) the pons, and the medulla oblongata. The cerebellum is involved in motor control

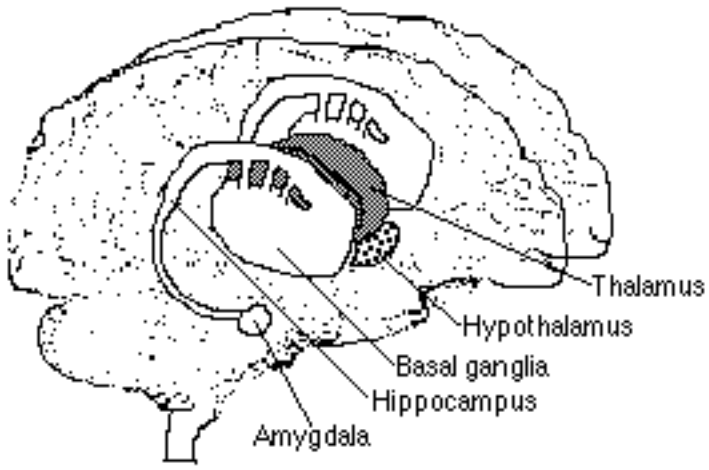


Figure 3.2: **Major Inner Brain Structures** (from [Carlson, 1991] pp. 89-90)

and coordination, and plays a part in some kinds of learning. It is believed that automatized skills like guitar playing are controlled by the cerebellum. The hindbrain is the oldest part of the brain, and is also primarily concerned with "lower" functions.

Each of these larger parts of the brain conceal inner systems of specific neurons, sometimes distinguishable by histological techniques, sometimes by pharmacological techniques, sometimes by electrical techniques or imaging. In all cases the basic substructure is composed of neurons, specialized cells that use electrochemical signals to communicate with each other and ultimately to control the behavior of the body.

### Major Neuron Structures

A typical neuron has four basic functional parts: dendrites, a cell body (soma), an axon, and terminal buttons (see 3.3). Each of these parts have complex structures, but their basic functional role can be easily understood without too much detail.

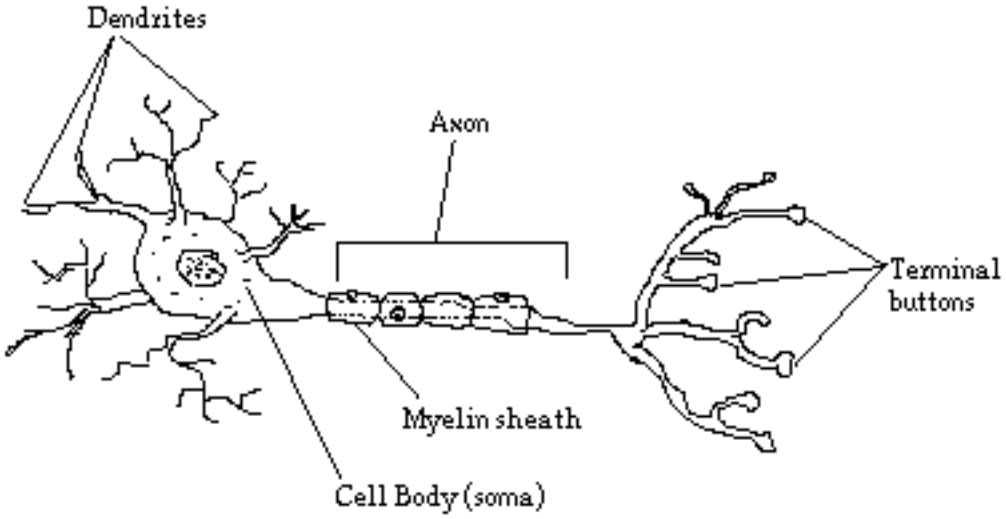


Figure 3.3: **Major Neuron Structures** (from [Carlson, 1991] p. 23)

Dendrites are the input side of the neuron, and branch out from the soma in various different ways (I have shown only one in the figure). They synapse with other neurons' terminal buttons, receiving chemicals which change the electrical potential of the cell, in one of two directions (more positive, more negative) and over various different time periods. More detail will come later when I describe synapses.

The soma of a neuron contains much of the life-sustaining structure of the neuron, including a nucleus, mitochondria, ribosomes, etc. It also synapses with other neurons. A single neuron may receive thousands of inputs from other cells, through its dendrites and soma. The axon is part of the output side of the neuron. When the electrical potential of the cell has been changed enough an action potential is generated, an electrical signal that travels down the axon to the terminal buttons. The terminal buttons have synapses at their ends, and they connect to the dendrites or soma of another neuron. When stimulated by an action potential they release their specific neurotransmitters.

**Neurotransmission**

Neurotransmission is a complex and not completely understood process, involving many different substances and complex interactions. The basic theory of neurotransmission is, however, not difficult to understand.[3] There are three basic events: release, binding at the postsynaptic neuron, and deactivation of the neurotransmitter (NT) in one of several ways. There are also several processes which regulate each of these events. Neurotransmitter release happens when synaptic vesicles filled with NTs in the soma or the terminal button join with the presynaptic membrane and expel their contents into the synaptic cleft. The NTs then attach themselves to receptors on the postsynaptic neuron and cause changes in that cell, either exciting it (changing the electrical potential toward generating an action potential), inhibiting it (changing the electrical potential away from generating an action potential), or changing some process in the receiving cell more permanently (as in the case of steroids). The NTs are then deactivated either by enzymes which break down the NT or by reuptake into the presynaptic cell.

Each of these steps is regulated by different processes. Release can be regulated anywhere along the synthetic chain for the creation of the NT, by the addition of more of the precursor or enzymes which aid the process for up-regulation, or by limiting the supply of precursor substance or blocking the action of the enzymes which catalyze the process for down-regulation. The binding activity of the NTs can be upregulated by diminishing the enzymes which break it down, by making the presynaptic autoreceptors less sensitive (thereby down-regulating reuptake or up-regulating release), or by adding another receptor agonist.[4] Binding activity can be down-regulated by the opposite mechanisms or by the presence of a receptor

[3] 3.4 ought to help make these processes and structures more clear.

[4] An agonist is a substance which has the same effect as the NT in question, whereas an antagonist has the opposite effect.

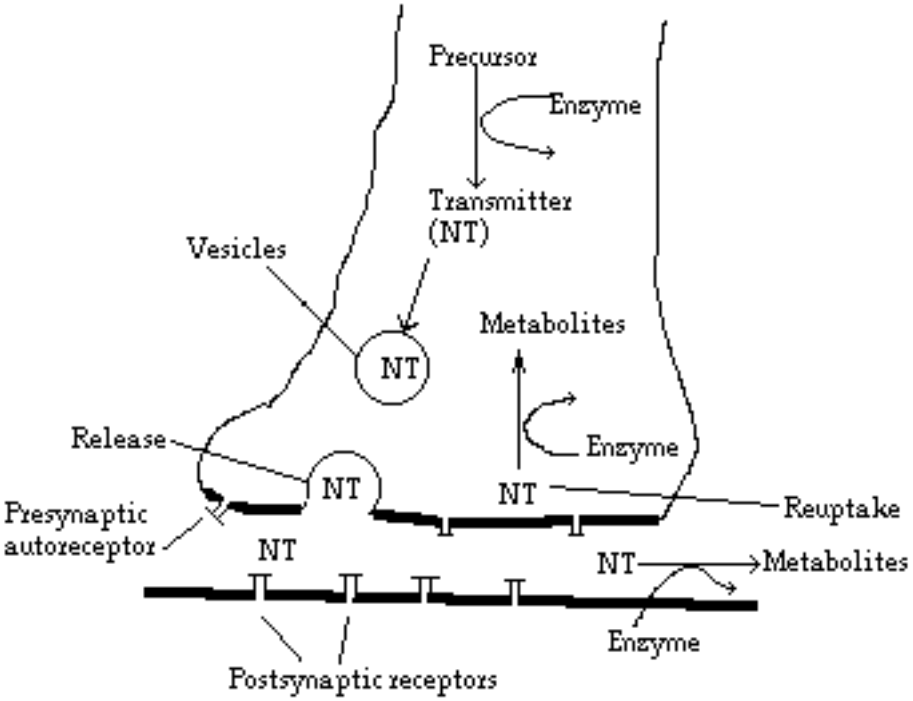


Figure 3.4: **Synapse Structure and Mechanism.** (from [Carlson, 1991], pp. 68-70)

antagonist. Deactivation can be regulated by many of these same mechanisms: changes in enzyme populations, presynaptic autoreceptor sensitivity, or reuptake activity.

### 3.1.2 Methods of Neuroscience

The brain as an organ is an amazingly complicated thing: composed of billions of neurons and attendant glial cells, supported by a specialized network of blood vessels and chemical factories, and thoroughly interconnected. Even in the interaction between one neuron and another there is sufficient complexity to frustrate any scientist: various neurotransmitters (NTs) and their synthesis, release and metabolisis mechanisms; the electro-chemical signaling mechanisms; complex multiple-NT receiving sites, etc. How to study this inscrutable object? Neuroscientists have developed a variety of techniques for teasing apart the complex structure of the brain, through histological, chemical, computer-aided and behavioral methods. Neuroscientific methods involve various ways of learning about the brain through stimulating, imaging, or destroying various systems of the brain to determine their physical structure and functional correlates. With this information, hypotheses about what each sub-part of the brain does can be made and tested. Stimulation, imaging, and lesioning can be accomplished via pharmacological (chemical), electrical, radiochemical, and computer-aided means, among others, and both within these techniques and between the broader methods neuroscientists use converging evidence to build up a model of the physical brain.

#### Pharmacological Techniques

Pharmacological investigative techniques involve the use of neurotransmitters and their antagonists, substances that resemble specific neurotransmitters (but have special properties), substances that do not resemble NTs but bind at their receptors

(i.e. psychotropic drugs), and substances that affect the metabolic chains which produce or break down neurotransmitters. Each of these three methods can be applied to the whole brain (non-invasive but often too broad)[5] or to small parts of the brain (through invasive microinjection.) The results obtained from these techniques can be maps of where specific neurotransmitters are used in the brain, maps of which NTs are bound at specific kinds of neurons, discoveries of new systems of receptor types, and behavioral changes brought about by pharmacological manipulation.

Another chemical technique uses the neurons' own system of axonal transport to reveal connections within the brain. If we want to trace all of the inputs to a specific brain area, we could inject into that area a stain which would be taken up into the axons projecting there and transported backwards along them, staining the cell bodies of those neurons. After dissecting and slicing the brain, these stained areas would make the connections visible.

An example of a substance that resembles a neurotransmitter but has special properties might be a radioactive analog which stays bound to a specific receptor type, like for dopamine. This analog would be administered, after which the subject would be killed and its brain would be sliced up. These brain slices would be used to expose film, so that the parts which were radioactive would show up. By overlaying this film on a photo of the brain slice, the researcher can identify where the dopamine receptors are in the brain. Using this technique, a global picture of dopamine receptors can be created, perhaps pointing to functional relationships between brain areas.

Pharmacological lesioning is another destructive chemical technique used to identify distinct receptor types or particular kinds of neurons. This method involves the administration of a toxic analog of a NT, like 6-hydroxy-dopamine, which binds to

[5]Administering drugs to the whole brain has been decribed as a "sledgehammer" technique, because anything given systemically will affect sometimes mutually antagonistic systems.

dopamine receptors on dopaminergic neurons and kills them. This is yet another way of developing a map of NT use in the brain,[6] but it can also be used with microinjection in behavioral studies. This enlarges the field of questions that can be answered: not only can we find out where NTs are used, created, etc., we can determine what behaviors or functions those NTs govern.

### Electrical Techniques

Lesioning can be done electrically as well as pharmacologically. Using a stereotaxic device[7] an electrode can be placed into a specific brain area. Running a current through the wires burns a very small portion of the brain, hopefully the proper brain area. This method has some problems, because it not only destroys the neurons in that area but any axons that pass through that area, so that the deficits or other behaviors observed could be from brain areas connected through the target area.

Another electrical technique involving the stereotaxic placement of wires in the brain is for recording, especially from single cells or very small groups of cells. This technique can yield very exact results, and distinguish very closely between different groups of neurons. Some important findings from this kind of methodology have been in the visual system, where different groups of neurons have been found to respond to moving stimuli, different colored stimuli, etc. Single-neuron recordings have even individuated between small areas of the visual processing system that distinguish between things moving vertically, horizontally and diagonally.

Event-related potentials (ERPs) are averaged recordings of the activity of many neurons. Electrodes are placed on the scalp which detect the small electrical potential changes caused by the firing of many neurons in the brain. Many recordings are

[6]Since many of these histological techniques are difficult and often depend on subjective judgement, converging evidence from different techniques can be crucial.

[7]A device for placing an electrode in an exact spot in the brain, basically a calibrated restraining device.

taken of the same kind of trial, and the information from all of them is averaged and analyzed. When this information is time-locked to some kind of stimulus or task, the relative changes in evoked potentials can be correlated with them. Different ERPs occur with different tasks, i.e. different ERP data between auditory processing and visual processing tasks.

### Computer-Aided Imaging

Some relatively recent advances in technology have greatly aided neuroscientists in their quest to understand the structure and functions of the brain[8] There are a few different techniques, each of which has its advantages and disadvantages, especially in resolution of time or detail; but between which we can once again use convergent evidence to test hypotheses.

Computerized tomography (CT) is an advanced x-ray technique. X-rays are sent through a patient's brain with a moving emitter-detector system, so that the brain can be scanned from all directions. This produces a very detailed two-dimensional x-ray of the brain, often used to detect tumors or to map functional deficits to physical deficits (for instance, in the case of a stroke on could find out which brain areas were damaged and correlate this with mental deficits). CT scans are limited to horizontal sections and do not show metabolic changes over time.

Magnetic resonance imaging (MRI) uses a strong magnetic field to force atomic nuclei in hydrogen in the brain to align their spins, and then hits them with radio waves. Different molecules emit different radio frequencies, which are detected and analyzed to create a picture of a slice of the brain, at any angle. Compilation of MRI images can be used to make a three-dimensional picture of an individual brain without invasive surgery, and are used for the same kinds of research as CT scans.

[8]The discussions of CT scans, MRI, PET scans and regional cerebral blood flow are adapted from [Carlson, 1991], pp. 113-116; and from [Churchland, 1993], pp. 217-221.

Positron emission tomography (PET) can be used to determine which brain areas are more active than others over time by their rate of glucose use. This is done by injecting the patient with 2-deoxyglucose (2-DG), a radioactive analog of glucose which emits positrons when struck by x-rays. The brain areas that are most active take up the most 2-DG, and so emit more positrons. This technique can be used to determine which brain areas perform which functions, i.e. flexing a muscle or doing math problems or talking. If this technique had better time and space resolution it could provide some of the most important data necessary for understanding the structure and function of the brain, but even as it is it has confirmed many results and hypotheses originally from other methods.

The other main computer-aided imaging technique for brain function is measuring the rate of cerebral blood flow (rCBF). This technique rests on the assumption that those brain areas which are most active will have higher rates of blood flow than other areas. The procedure involves having the subject inhale radioactive gas (xenon 133) and perform a task while surrounded by very sensitive detectors, which map the rCBF based on how radioactive each brain area is. Thus there is another measure which can determine which physiological areas are used for which cognitive (or motor, or whatever) tasks.

### 3.1.3 Neuroscientific Theories of Consciousness

In 1990 Francis Crick[9] and Christof Koch published a paper [Crick and Koch, 1990] called "Toward a neurobiological theory of consciousness." This paper explores the "binding problem": How does the brain bind together all of the different subcomponents of sensory processing into one coherent percept? They narrow their scope to vision, because so much is known about the visual system, and ask more specific questions like "How does the brain bind the shape of an object, its color, its mo-

[9] A scientific giant, of DNA fame (with Watson.)

tion, and its distance into the perception of a single object?" All of these separate qualities of a visual perception are known to be processed in different areas of the brain, and yet we do not perceive them separately. How are they bound together? Is this binding mechanism an integral part of consciousness? Crick and Koch used findings from neuroscience to guide their theorizing on these questions, and have developed an exciting theory of consciousness at the neuronal level.

As well as describing how the brain might bind the inputs from various cortical areas together, Crick and Koch speculate on short-term memory and attention, two of the main ingredients in conscious perception. Their speculations are unfortunately not very detailed, and reveal a large gap in our knowledge of neuroscience.[10] Short-term memory is "likely to be distributed throughout the appropriate cortical areas" ([Crick and Koch, 1990], p. 270) with the particular events remembered being stored in the relevant cortical areas: visual memories in the visual cortex, auditory memories in the auditory cortex, etc. The neuronal realization of short-term memory could work in one of three ways: "(i) the strength of certain synapses is temporarily increased or decreased; (ii) a neuron keeps on firing for some time due mainly to its intrinsic biophysical properties; or (iii) a neuron keeps on firing mainly due to extrinsic circuit properties ('reverberatory circuits')." ([Crick and Koch, 1990], p. 270)

Their speculation for how the visual system decides where to focus attention is to suggest:

> ...some sort of topographic saliency map that codes for the conspicuousness of locations in the visual field in terms of generalized center-surround operations. This map would derive its input from the individ-

[10]This gap might exist because neuroscientists do not know what sort of phenomena to look for: both of these processes are dynamic rather than static, and so can't be investigated by the sophisticated histological techniques available; and even if they could see into the brain while it was working it would be hard to pick out the details in question.

> ual feature maps and give a very 'biased' view of the visual environment, emphasizing locations where objects differed in some perceptual dimension, i.e. color, motion, depth, from objects at neighboring locations. ([Crick and Koch, 1990], 1990, p. 271)

In other words, the way the neurons are connected together would cause some kinds of features to 'pop out' from the background. Something like this could even explain how personally important objects could be more immediately recognized. This is not a wild notion, as some sort of 'memory' might easily be built into neurons in this system which 'remembers' which things are important via stronger synaptic connections between the relevant neurons, or via 'top-down'[11] influence from higher level systems.

Once the attentional selection has taken place, there must be some way to bind together these disparate groups of neurons so that they form a coherent perception, and then a mechanism to put them into short-term (and eventually long-term) memory. Crick and Koch have come up with a very plausible theory about how this is done: once the object in question has started all the relevant groups of neurons firing some other mechanism causes them to continue firing at a specific frequency, 40 Hertz.[12] This would be accomplished through some sort of central feedback mechanism. Objects whose sensory characteristics are thus bound are then placed in working memory, and perhaps it is these oscillations which activate the mechanisms of working memory. Another possible consequence of this theory is that the 40 Hz oscillations explain the limited nature of attention: since only a few separate oscillations can be occurring simultaneously, only a few distinct things can be paid attention to at the same time ([Crick and Koch, 1990], p. 273).

[11] "Top-down" influence is a cognitive notion, which means that important things at higher (conscious or unconscious) levels influence the actions of very low-level systems, speeding their responses to those things.

[12] 40 Hertz = 40 times per second.

Crick and Koch have come up with a very plausible mechanism for explaining how the visual system binds together the neurons involved in seeing, as well as a possible model for other sensory modalities. Their model also addresses attention and short-term memory, although some questions remain: How does the central feedback mechanism work? How does it choose when and which neurons should be bound together? Which of their postulated mechanisms for short-term memory is right? The explanation they give for attention, a "winner-take-all process" is quite reasonable, and fits with other theories of low-level information processing.[13] Fortunately most of their speculations are empirically testable. Unfortunately, however, they are resolute type-physicalists: "...We believe that the problem of consciousness can, in the long run, be solved only by explanations at the neural level." ([Crick and Koch, 1990], p. 263) They do not seem to really believe this, as at the end of their conclusion they speculate that the reason consciousness is so mysterious is that we cannot conceptualize such a massively parallel system in operation, that the ability of the brain to handle a huge amount of information all at once is what makes it so mysterious. This is not a neural-level explanation, but a system level one, and their theory is also a system level explanation even if they don't know it.[14] If it was a neural-level explanation, it would consist of lists of which particular neurons were involved, or at least which subsets of neurons, and it would not be tenable if other brains worked with slightly different (or very different) sets of neurons. The brain is a flexible machine, capable of "rewiring" itself, and so a system-level explanation will succeed over a neural-level explanation.

[13]Specifically, the "pandemonium" model, in which the neurons which "shout the loudest" (have the strongest connections or are pre-primed) are those that pass along the results of their processing to the next higher level. This theory is also discussed in cognitive science models of information processing.

[14]Here is an example of why philosophy can help scientists: they are ignorant about the positions they themselves hold, and then develop prejudices based on their (false) conceptions, and so productive cross-fertilization with other fields is stymied.

Although at one level neurons are like the binary parts of a computer in that they either fire (generate an action potential) or do not, it should be clear that the actual processes which culminate in firing are quite complex and flexible, even without all of the details.[15] Even the most sophisticated attempts at modeling brain systems don't come near the richness of actual neuronal interaction. It should also be clear that this level of explanation will not produce a theory of consciousness: even though I believe that these processes are what consciousness comes from fundamentally, any explanation which gives answers consisting of lists of neuron-level details will be unsatisfying. An example from physics will serve here: If I want to know how my favorite mug got broken, I am looking for an answer like "Joe dropped it in the sink" rather than a list of all of the molecular-level interactions that occurred. To understand how the brain produces consciousness we must look at a higher level, understanding all the while what is going on at the fundamental level. Cognitive neuropsychology is an attempt at this sort of explanation.

## 3.2 Cognitive Neuropsychology

Cognitive neuropsychology[16] is the branch of psychology that tries to map psychological findings onto neurophysiological findings. It is an attempt to integrate our understanding of the functions of the brain with our understanding of the structures of the brain. Their research strategy often uses humans who have various kinds of brain damage or have had intentional brain surgery[17] and determines what kinds of deficits result from what kinds of injuries, which kinds of losses can be recovered and

[15] I have omitted some of the richness of synaptic activity in the interests of brevity and clarity. There is much more to know about each of the subparts I have described. [Carlson, 1991] is a reasonable introductory text for this information, as well as P. S. Churchland's Neurophilosophy [Churchland, 1993].

[16] Hereafter "neuropsychology" refers to cognitive neuropsychology, rather than neurology or any other form of neuropsychology.

[17] Hence, it is often called colloquially "hole-in-the-head" research.

which kinds can't, and thereby which physical areas of the brain serve which mental functions. In this section I will describe some of the findings of neuropsychology as well as the methods used to obtain this information.

Neuropsychologists study, for the most part, people with brain damages of various kinds and degrees, and have therefore developed a system of categories to use in describing these disorders, including agnosias, aphasias and apraxias. Agnosias are disorders in sensory processing and the organization of sensory information. These are caused by lesions in cortical association areas. An example is visual agnosia, in which the person affected can see, but doesn't recognize what they see. Aphasias are language disorders caused by lesions in specific areas of the cortex, including Broca's area and Wernicke's area. Lesions to Broca's area cause expressive aphasias, while lesions to Wernicke's area cause receptive aphasias.[18] Apraxias are disorders in the planning and execution of movements, and are often caused by damage to the frontal lobes of the cortex[19][Gleitman, 1986], pp. A30-31). All of these disorders can also result from lesions in different areas, which is a problem neuropsychological explanations of brain function must take into account. The basic difficulty for neuropsychologists, mapping lesion damage to cognitive deficits, has gotten much easier due to advances in technology. Where Broca or Wernicke had to wait until their patients died before examining their brains for lesions, current neuroscientists can use MRI, PET and CT scans to study people's brains while they are still alive. These techniques also have helped characterize degenerative disorders, because the damage can be tracked as it progresses instead of only at the end of the disease after

[18]This is a gross oversimplification. Cognitive neuropsychologists have also believed that Broca's and Wernicke's areas control syntax and semantics, respectively, not to mention the fact that patients with damage in either of these areas are likely to show deficits in both comprehension and production of language. Nevertheless, this simple deficit-lesion mapping is roughly true.

[19]Where much planning probably takes place, possibly all planning. (There will be more on this subject in the section on problem solving, and again in the section on neuropsychological theories of consciousness.)

the patient has died.

Neuropsychology's unique combination of cognitive psychology and neurology allows a kind of theorizing impossible in the two separate areas: cognitive psychologists' models don't refer to physiological mechanisms, while neuroscientists' models don't refer to cognitive functions. Neuropsychologists are crippled, however, by their inability to perform controlled experiments on humans, so the methods they use are inherently open to criticism. Nevertheless, some of the most exciting things we have learned about the brain have come from this field.

### 3.2.1 Findings of Neuropsychology

People who have experienced strokes, gunshot (or other) wounds, suffocation, or other traumas which either selectively or generally damage the brain have been the traditional subjects for neuropsychologists. The specific kinds of disorders they have identified have been fairly successfully mapped to different brain areas, providing an excellent functional/physical map of the brain. From these studies and more recent computer-aided studies on normal subjects we know what certain areas of the brain do, what they are for, and how they are organized. What follows[20] is a brief catalogue of brain areas and their relevant functional and hierarchical connections, from sensory input on up to conscious planning. The broad areas involved are derived from a breakdown of cognitive abilities: visual (and other sensory) processes and attention; language comprehension, production, and reading: short-term and long-term memory; and problem solving ([McCarthy and Warrington, 1990]). Within each of these domains smaller cognitive components may be selectively damaged—much of the success of this field is due to people with very minor brain damage which impairs only certain of their cognitive abilities and not others.

[20]This section draws heavily from [McCarthy and Warrington, 1990], chapters 2-4, 6, 8-10, and 12-16. Most of the specific information is relatively uncontroversial, as reference to other texts (e.g. [K. and Valenstein, 1985]) will show. Terms in boldface are referenced in table 3.1

**Visual deficits** fall under the category of **agnosias**, and involve various kinds of "mindblindness." ([McCarthy and Warrington, 1990] pp. 22-55) These disorders may be divided into three distinct kinds of deficits:

1. Deficits affecting sensory processing,

2. Deficits affecting the perceptual analysis of known objects (apper-ceptive agnosia), and

3. Disorders in deriving the meanings of objects (associative agnosia).

These disorders involve different aspects of visual processing each with its own specific constellation of deficits. **Visual processing** deficits can show up in discrimination tasks involving acuity,[21] shape discrimination, and color discrimination. **Apperceptive agnosias** cause deficits in tasks involving recognition of objects from incomplete drawings, in different lighting conditions, from different views, and from overlapping objects. **Associative agnosias** cause difficulty in understanding the functions and meanings of objects, so that things like cups and forks can be recognized but not used, or identified as normal or abnormal but not named. Each of these disorders comes from brain damage to different areas. Visual processing is degraded by damage to the occipital lobe, and is often unilateral (and so causes half the visual field to be degraded.) Apperceptive agnosias arise from damage to the right hemisphere, in the parietal lobe. Associative agnosias arise from damage to the left occipitotemporal region, at the junction of the occipital and temporal lobes. These three different families of deficits are believed to correspond to three separate stages of processing visual information. The first stage is visual sensory processing, where low-level features are distinguished. The next stage is the deriva-

[21] Acuity includes the ability to detect presence or absence of light, number of objects perceived, changes in contrast, and target resolution.

tion of a structural description of the object, and the last stage assigns meaning to the percept.

**Language comprehension** of course involves auditory processing, as well as final comprehension of word meanings. Even with normal hearing, however, some people show deficits in word comprehension. At a very low level some patients have trouble **discriminating phonemes**, the basic auditory components of spoken language (for example, distinguishing 'ba' from 'pa'). These deficits show up in people with unilateral left temporal lobe lesions. At a higher level are subjects who have no trouble distinguishing phonemes but for whom **single words** don't make sense. Often these disorders affect only infrequently-used words, but not always. There are even cases where patients lose the ability to comprehend only certain **categories of words**, like animals or body parts. These deficits also seem to be caused by left hemisphere temporal lobe lesions, but some of the more specific deficits come from other areas, like color-word comprehension impairments caused by left occipital lobe lesions. Here again is evidence for stages in processing of words: first, temporal resolution, then phonemic categorization, and finally word meaning.

**Sentence comprehension** and **construction** may also be selectively damaged. Patients with the former kind of deficit cannot understand instructions like "The leopard was killed by the lion."[22] They can understand the words, but cannot say which animal was killed—this indicates some sort of syntactic deficit rather than a semantic one. Patients with the latter kind of disability show a variety of different deficits in producing sentences. These deficits often show up at the level of semantics. Sentences come out jumbled or with many repetitions of common words, or in the most extreme cases (jargon aphasia) can even be complete nonsense, sometimes called "word salad."[23] Sentence comprehension deficits seem to arise from lesions

[22]This example is from Enriqueta Canseco-Gonzalez, personal conversation March 1994 (Thanks!)

[23]An excellent example from [McCarthy and Warrington, 1990], p. 182: "...They were in-

almost anywhere in the left hemisphere, while sentence construction deficits seem to arise from lesions in the anterior left hemisphere, especially in Broca's area.

There are two main types of **reading deficit**, both called **dyslexias**. Those "affecting the visual processing of words are called **peripheral** or **visual word form dyslexias**; those affecting the ability to derive sound or meaning from print" are called **central dyslexias** ([McCarthy and Warrington, 1990], p. 215). For example, someone with visual word form dyslexia could not even copy a written word, while someone with central dyslexia could copy the word but not read it or understand it. The first kinds of dyslexia include problems with letter order within words and neglecting the left or right halves of words (**neglect dyslexia**). **Central dyslexias** called surface dyslexias cause people to mispronounce words in systematic ways, as if they had forgotten some of the rules of pronunciation, e.g. pronouncing "lace" as "lake" ([McCarthy and Warrington, 1990], p. 221). Another kind of central dyslexia is called **phonological dyslexia**, in which the patient cannot sound out words but uses a 'sight vocabulary' so that nonsense words and uncommon words cannot be read but common words can. The kinds of brain damage which produce visual form dyslexias are in the occipitotemporal region. Neglect dyslexias occur with brain damage to the contralateral side of the brain from the deficit, so that left-half dyslexia occurs with right-side lesions. Surface dyslexia is associated with posterior left temporal lobe damage, while phonological dyslexia is associated with temporoparietal lesions.

Within the category of **short-term memory** (STM) are two kinds of deficits: those which restrict the capacity of memory, the number of items which can be remembered exactly, and those which restrict the duration of short-term memory. In turn, each of these deficits is different depending on what modality the task is in. **Auditory-verbal STM** deficits are those in which patients have impaired "ability

vernted kassterz wiss kisstek an niches ik hampess for nekstes an terress and so on."

to repeat spoken lists of words, numbers and syllables" ([McCarthy and Warrington, 1990], p. 276). **Visual-verbal STM** deficits impair the ability to remember items presented pictorially. **Visual-spatial STM** deficits impair the ability to remember spatial features of stimuli, like matching groupings of dots with a previously shown stimulus or remembering a sequence of touches on blocks set in a random pattern. Auditory-verbal STM deficits arise from left inferior parietal lobe lesions, visual-verbal STM deficits arise from left occipital lobe lesions, and visual-spatial STM deficits come from lesions in the right occipitoparietal region ([McCarthy and Warrington, 1990], p. 282-284).

Disorders of **long-term memory (LTM)** can be divided into two main types: those involving a person's memory of their life (**autobiographical memory**) and those involving general knowledge (**material-specific memory**). **Retrograde** and **anterograde amnesias** are the main disorders of the first type. **Retrograde amnesia** is the inability to remember events that occurred before the onset of the disorder. People with this amnesia remember general knowledge and salient facts about themselves but few details: their names and usually their marital status are remembered, but what schools they attended, what their occupation is, and other details are lost. **Anterograde amnesia** is the inability to remember anything that has occurred since the onset of the disorder. This inability to lay down new permanent memories can be quite debilitating, especially as it is often accompanied by retrograde amnesia. Some types of learning are possible, but they are not explicitly recognized. For example, an amnesic might learn a piece of music gradually over several days, and be able to play it very well after a time, but still insist that they had never seen the piece of music before. Brain damage which affects the limbic system, including the hippocampus, amygdala, thalamus and mammilary bodies can cause global amnesias ([McCarthy and Warrington, 1990], pp. 296-315).

Material-specific long term memory can be dissociated into **verbal** and **non-**

**verbal** memories. **Verbal memory** deficits are different from aphasias—they are the inability to remember sentences, short stories, and paired-associate word lists. **Nonverbal** deficits impair the ability to remember abstract two-dimensional pictures, to recognize recurring nonsense figures, to remember how to get through a maze (spatial memory), and to remember faces. The brain damage which causes these deficits is usually unilateral, with left-side lesions damaging verbal memory and right-side lesions damaging visual memory. These lesions can be in the frontal, temporal or parietal lobes, suggesting breakdowns at many possible levels.[24]

The last kinds of deficits I will discuss in this section are those in the broad category of **problem solving**. This includes:

1. focused attention,
2. higher-order inferences,
3. formulation of strategies,
4. flexibility, and
5. evaluation of the outcome ([McCarthy and Warrington, 1990], p. 344).

**Focused attention** deficits are those in which the ability to ignore distractions and focus is lost, so it is impossible to concentrate on a task. Focused attention may in turn be broken down into sustained attention and selective attention. Sustained attention deficits show up as the inability to ignore distractions, while selective attention deficits show up as inability to avoid producing the "most automatic or habitual" response ([McCarthy and Warrington, 1990], p. 344-345). Disorders in **higher-order inference** production are manifested in the inability to form abstract concepts about things. "For example, although a saucepan

[24]Moscovitch & Umilta discuss some of the theoretical implications of memory failures in the first chapter of [Schwartz, 1990].

and a pair of scissors can differ ... [in many ways] they are both instances of the higher-order or more abstract classes of 'metal things,' 'household objects,' ... [etc.]" ([McCarthy and Warrington, 1990], p. 346). Disabilities in the **formulation of strategies** impair the production of suitable plans of action for the task at hand. Sometimes these disorders show up as perseveration, where the same thing is tried over and over again, or as failure to look ahead, to allow for intermediate steps to a goal. **Flexibility** deficits are those which prevent the adoption of different perspectives, as in matching left and right hands with a drawing of another person. The ability to **evaluate outcomes** may also be impaired, either because of inability to use feedback or inability to "solve a problem 'according to the rules'" (p. 351). Perseveration shows up here too, in the inability to stop doing something even when told it is wrong.[25] All of these problem solving deficits have been attributed to frontal lobe damage. Luria[26] hypothesized that attentional deficits come from medial frontal lobe damage, and Milner[27] argues that dorsolateral damage causes some of the other cognitive deficits.

As table 3.1 shows, quite a large amount of the functional architecture of the brain has been mapped. Even the brief account I have given here[28] contains enough information to drive theorizing about cognition. How reliable is this information? How have neuropsychologists studied these areas? The next section will deal with these questions, before turning to the theories this information has inspired.

[25] As in the inability to switch sorting criteria in the Wisconsin Card Sorting task (p. 351-352) which is comprised of cards which can be sorted according to color or shape or number of items.

[26] Cited in [McCarthy and Warrington, 1990], p. 357.

[27] Cited in [McCarthy and Warrington, 1990], p. 357.

[28] There is much more, especially dealing with lower brain areas, which would not fit here. For more detail see [McCarthy and Warrington, 1990]; [K. and Valenstein, 1985]; and [Carlson, 1991].

| **Cognitive Deficits and Associated Brain Damage** | |
|---|---|
| Visual deficits | Occipital |
| Agnosias | Occipital, parietal, temporal |
| Visual processing | Unilateral occipital |
| Apperceptive agnosias | Right parietal |
| Associative agnosias | Left occipitotemporal |
| Language comprehension | Left temporal |
| Discriminating phonemes | Left temporal |
| Single words | Left temporal |
| Categories of words | Left occipital |
| Sentence comprehension | Left hemisphere (general) |
| Sentence construction | Broca's area |
| Reading deficits (Dyslexias) | Occipitotemporal |
| Visual word form dyslexias | Occipitotemporal |
| Central dyslexias | Posterior left temporal |
| Neglect dyslexia | Unilateral damage, Right parietal |
| Phonological dyslexia | Temporoparietal |
| Auditory-verbal STM | Left inferior parietal |
| Visual-verbal STM | Left occipital |
| Visual-spatial STM | Right occipitoparietal |
| Autobiographical memory | Limbic system |
| Retrograde amnesia | Limbic system (hypothalamus, thalamus) |
| Anterograde amnesia | Limbic system (hippocampus) |
| Verbal memory | Left frontal, temporal, parietal |
| Nonverbal memory | Right frontal, temporal, parietal |
| Problem solving | Frontal lobes |
| Focused attention | Medial frontal |
| Higher-order inference | Dorsolateral frontal |
| Formulation of strategies | Dorsolateral frontal |
| Flexibility | Dorsolateral frontal |
| Evaluate outcomes | Dorsolateral frontal |

Table 3.1: Cognitive Deficits and Associated Brain Damage

### 3.2.2 Methods of Neuropsychology

Neuropsychologists study people with brain damage and try to find correlations between the kinds of cognitive deficits they show and the pathologies which cause them. Finding people with very localized damage and very few cognitive deficits is difficult: often someone with brain damage has extensive or spread out damage. Strokes and progressive degenerative disorders like Alzheimer's, for example, often affect many different brain areas or cause general damage. There are few cases of such specific disorders, but people with interesting and limited deficits have been studied for 100 years [Shalice, 1991]. Group studies have not often been performed, as deficits even within a single syndrome (like apraxias) can be very specific and individuals can differ widely from the mean, so that comparisons are useless. This has led many neuropsychologists (Badecker, Caramazza, Shallice) to look more to single case studies to learn the breadth of possible damages and by inferring cognitive processes from these create a map of brain function. Single case studies suffer from many disadvantages as well:

> Even relatively favourable critics like Zurif ... point out, however, that single case studies seem especially prone to problems from the selection of premorbidly atypical subjects and from the adoption of idiosyncratic strategies by individual patients. In addition, to extrapolate from either group or single case studies one must assume in practice that no critical reorganization of function takes place after the lesion; there is little or no evidence for this. ([Shalice, 1991], p. 430)

Shallice nonetheless believes that single-case studies can be valid, as long as their weaknesses are acknowledged. He goes on then to provide a set of assumptions and guidelines for inferring mental structure from neuropsychological findings, raising issues best discussed in the next section. We now have a picture of the difficulties

that face neuropsychologists in inferring anything from either group or single-case studies. How, then, do they determine which cognitive differences are qualitatively different enough to qualify as separate disorders?

The usual method for determining the extent of certain disorders is called dissociation. In dissociation, lesions of an area are associated with cognitive deficits and are dissociated from other lesion-deficit pairs. So a lesion to area A would cause deficit A', and a lesion to area B would cause deficit B'. If these correlations can be found consistently, and if cases exist where such fine-grained distinctions can be made, then these brain areas can be linked to the cognitive functions which are impaired when they are damaged. Even better than dissociation is double dissociation. In double dissociation, lesions to area X cause deficit X' and not deficit Z', while lesions to area Z cause deficit Z' and not deficit X'. This shows not only that the relevant areas are important for the relevant functions but also that they are solely responsible for those functions, because only damage to that particular area causes those particular functional deficits.

Tests of cognitive functions are, unsurprisingly, often the same kinds of tests used by cognitive psychologists. The main difference is that neuropsychologists are looking for deficits from normal data rather than trying to find what the normal data is. While neuropsychologists may claim (fairly) to have sparked the implicit/explicit theories of cognitive psychology with the studies of amnesic patients, many of the tests of these functions have been developed in cognitive psychology. These include methods of probing for information that people fail to recall (explicit) by asking questions that force them to guess or to make choices based on "preference" (implicit).

### 3.2.3 Neuropsychological Theories of Consciousness

In 1983 Jerry Fodor resurrected the faculty psychology theory of mind in his book The Modularity of Mind [Fodor, 1983]. Faculty psychology means "the view that many fundamentally different kinds of psychological mechanisms must be postulated in order to explain the facts of mental life" (p. 1). The current version of this theory is now called the modularity theory, because it is based on the idea of separate modules which perform different cognitive functions. The modularity thesis is a marriage of cognitive psychology and the evidence from neuropsychology: cognitive psychology provides some regularities and theories of mental processes as well as ideas about information processing, and the evidence from neuropsychology provides a solid grounding for some of these distinctions.

What are modules? In Fodor's theory they are informationally encapsulated, domain-specific information processors. Informational encapsulation means that higher-level thoughts and information cannot affect their functioning: modules are discrete mechanistic processors, which take in information, process it, and send the product of their computations onward, without interference from above. Furthermore, higher levels cannot access the items processed by modules: the original data is lost, and cannot be accessed; only the results (categorizations) of the module's processing are available to higher levels. Domain specificity refers to the unique functions of each kind of module. For example, an auditory module would handle only sound information, the transformation of sound waves into neural impulses and some of the primary processing necessary for pre-semantic understanding of sound, and would not be used for any other domain. If a specific module is destroyed or damaged, only that function would be lost, so if an auditory module was destroyed deafness would result, but not the ability to understand written words. Information processing by modules is understood as a mechanistic process, probably performed

via a parallel-processing architecture rather than a serial processing architecture.[29] These characterizations work well for sensory processing, for the input side of consciousness, but what about the "higher levels"?

Input modules are fast, informationally encapsulated, have "shallow"[30] outputs, and are associated with fixed neural architecture (which accounts for the very specific deficits observed by neuropsychologists ([Fodor, 1983], pp. 61-99)). To explain how the output of these modules is understood and combined into a representation of the world and how humans use this information to think and guide behavior, Fodor posits another kind of structure: central systems. Central systems are slow, unencapsulated, deep, global, voluntary, widely neurally distributed, and have multidirectional information flow rather than one-way input-output information flow. Central systems are not modular for these reasons, according to Fodor, and he says there is no evidence for or against their modularity.

What central systems do is "fixation of belief (perceptual or otherwise) by nondemonstrative inference" ([Fodor, 1983], p. 104). In other words, they test hypotheses about the world (questions brought up by information from the modules) based on everything that is known. These processes are probably largely unconscious, so Fodor makes an analogy with scientific discovery to explain how they test hypotheses. Analogical reasoning seems to be the way many scientific discoveries have come about: Someone makes a connection between a regularity in one domain, like stock market dynamics, and another, like natural selection ([Fodor, 1983], p. 107). This kind of reasoning requires access to many different domains; in principle to all

[29] A discussion of information processing, serial vs. parallel processing, pandemonium models based on feature-detectors, etc. is in the section on cognitive psychology.

[30] Fodor's distinction between "shallow" and "deep" outputs involves the availability of these outputs to other systems, like central systems, and is part of the philosophical dispute about where data processing turns into content-bearing representations. Needless to say, this is too thorny a topic to go into here. Suffice it to say that "shallow" output never gets to consciousness—it is like raw data that we never see. (See the subsection on categorical perception in the Cognitive Psychology section.)

domains. With these two kinds of systems linked in hierarchical order, Fodor has created a model of mind, with modules processing input information, presenting the results to central systems, which test hypotheses about these inputs based on memory and form beliefs and cause actions.

Tim Shallice [Shalice, 1991] has trouble with Fodor's central systems. He cites neuropsychological evidence which shows that certain entire domains of knowledge which should be part of distributed central systems can be selectively damaged, as in acalculia ([Shalice, 1991], p. 436). Furthermore, some thought processes which are paradigmatically central can be selectively affected, in different frontal syndromes, like planning or flexibility disorders (see table 3.1). To explain these syndromes, Shallice and Norman (1986)[31] posited two new mechanisms, a contention scheduling process and a supervisory system. The contention scheduling process selects between routine operations. This routine selection of routine operations is decentralized, and the operations themselves are thought of as a large set of programs or schemas. These thought or action schemas are in competition, and are selected when their level of activation (caused by lower-level modules) rises above a threshold, and are then carried out. The supervisory system works above the contention scheduling system, activating or inhibiting different schemas. This system can learn from mistakes, deal with novelty (situations for which there is no pre-existing schema), make decisions, and inhibit impulsive schemas[32]([Shalice, 1991], p. 436). Damage to the supervisory system causes over-reliance on contention scheduling, which can cause the kind of perseverative errors shown by frontal syndrome patients.[33] Shallice has one more interesting detail to his theory—he suggests that the purpose of episodic

[31] In [Davidson et al., 1986].

[32] Willpower, or the ability to put off short-term benefits for greater long-term benefits.

[33] Tiredness or distraction could also cause these kinds of errors, called action slips. For example, when driving a familiar route (to school, say), one might be inteding to go someplace entirely different but end up at school accidentally: The activated schema was not inhibited at the right time by the supervisory system.

memory (autobiographical memory) may be to provide the supervisory system with some way of approaching a new problem when no schema is available or strongly triggered by the current environment or goals. An analogous episode may be retrieved containing a solution or the means to a solution. Shallice, by expanding on Fodor's ideas of modularity, has created a model of mind which can be quite successful in explaining conscious phenomena and is based on strong empirical evidence from neuropsychology.

The last neuropsychological theory I will present is from Moscovitch and Umilta, 1989 (in [Schwartz, 1990]). This model argues with many of Fodor's criteria for distinguishing modules. They argue that "deployment of attention, which can be mediated by a central process, can be both mandatory and rapid" ([Schwartz, 1990], p. 4), as in the 'cocktail party' effect: it is impossible not to attend to one's name being spoken even when one's attention is focused elsewhere. Therefore, speed may not be an essential characteristic of modules. Three of Fodor's other criteria are also nonessential: "association with a fixed neural architecture, manifestation of characteristic and specific breakdown patterns, and a characteristic pace and sequencing during development" ([Schwartz, 1990], p. 4). Fixed neural architecture may be found for central processes (like memory and attention), such that focal brain damage can impair them selectively. Memory and attention also have characteristic breakdown patterns and regular developmental sequences. Since these criteria apply equally well to modules and central processes, something must be wrong with Fodor's theory.

Moscovitch and Umilta go on to criticize some of what they think are the essential characteristics of modules and to provide ways of operationalizing these criteria in neuropsychological terms. These essential characteristics are domain specificity, information encapsulation, shallow output, and inaccessibility of intermediate-level representations (nonassembly). Domain specificity is problematic because defining

domains is so difficult: the best definitions seem to come from people with specific kinds of deficits, but their damage is rarely specific enough for a strict definition. Information encapsulation may be better defined by a combination of neuropsychological criteria: to establish that a modular process is informationally encapsulated, double dissociation evidence from patients with focal brain damage and sparing of function evidence in patients with degenerative dementia may both be needed ([Schwartz, 1990], p. 8). Shallow output can be demonstrated using explicit and implicit tests. In an explicit matching for identity task, a demented patient could perform well in spite of not being conscious of what the objects were ([Schwartz, 1990], p. 9). In an implicit recognition task, patients with prosopagnosia[34] showed higher skin-conductance response to familiar faces even though they denied recognizing them. Intermediate-level representations should not be consciously accessible at all. However, normal subjects show sensitivity to intermediate-level modular information: even though people are not aware of subphonetic differences, "reaction times in a phoneme-matching task are influenced by them" ([Schwartz, 1990], p. 11). Brain damage can reveal some of these pre-output steps as well, as in patients who have visual system damage "report seeing a smoothly moving object as a set of static pictures located at various points in the trajectory, much like superimposed stop-action, stroboscopic pictures of a moving object" ([Schwartz, 1990], p. 12).

From the neuropsychological evidence presented in their article, Moscovitch and Umilta draw the conclusion that modules can be assembled. A good example of an assembled module might be reading, which is slow and difficult to learn, and requires central process-type thinking to learn, but eventually becomes completely automatized and displays many of the characteristics of modules ([Schwartz, 1990], p. 12). Although they have found much to criticize in Fodor's criteria for modularity, they still wish to retain the idea of modules. The reason behind this is that they

[34]The inability to recognize faces.

believe

> ...that to present to the central processes veridical information about the world quickly, efficiently, and without distortion from the beliefs, motivations, and expectations of the organism, something like modules that are immune to higher-order influences must exist. ([Schwartz, 1990], p. 13)

To save the useful idea of modularity and at the same time account for its deficits, three types of modules are posited, along with the idea that modules (by their definition) can be assembled. These three types of modules are: Type I, basic modules (like Fodor's); type II, innately organized modules; and type III, experientially assembled modules. Type I modules deal with relevant and predictable environmental stimuli. This includes basic sensory perception in each sensory modality, i.e. colors, acoustic frequency, sound location, visual location, depth, faces, and perhaps emotions.[35] Type II modules are innate organizations of basic modules "whose output is integrated or synthesized by a devoted, nonmodular processor." ([Schwartz, 1990], p. 15) Devoted means that these higher-order modules can only work with their specific basic modules. An example is vision, which integrates the output from modules like motion detectors, color detectors, shape detectors, etc. into a coherent percept.[36] Type III modules are created by central processes which assemble type I and type II modules together to create functions which become modular with practice. ([Schwartz, 1990], p. 17) Examples of type III modules are reading and bicycling, as opposed to type II modules which govern speaking and walking. Both types must develop, but type II modules have innately specified development whereas type III modules are consciously constructed.

[35]These last two may be handled by type II modules, although the neuropsychological evidence suggests that they satisfy all the criteria for basic modules.

[36]This sounds much like Crick & Kochs' [Crick and Koch, 1990] binding problem—perhaps their solution for consciousness is merely a solution for the action of type II modules!

Central processes still must be explained. Moscovitch and Umilta propose to explain them by virtue of their functions, rather than the kind of information they deal with, as in modules. They should still show specific deficits following damage, but the deficits would be described as a loss of function rather than a loss of knowledge. Central systems may also be neurologically defined by their connections: "input pathways to modules should be fewer than those to central systems." ([Schwartz, 1990], p. 20) The distinctions between central systems and modules can become difficult to define, as central systems may be selectively damaged, or the input pathways from modules to central systems may be damaged. When the input system is not damaged but the part of the central system that processes its output is damaged, is that a modular deficit or a central deficit? Because of these difficulties, Moscovitch and Umilta suggest "at the global level the prefrontal cortex behaves as a central system, but at the local level the prefrontal cortex (and other cortical areas) may resemble modules." ([Schwartz, 1990], p. 21)

Since both the inputs and information handled by central systems are by definition totally flexible, functional descriptions of central processes will be more useful. The four central processing functions described in this theory are "(1) forming type II modules, (2) forming type III modules, (3) relating information to general knowledge, and (4) planning." ([Schwartz, 1990], p. 28) Both (1) and (2) are learning functions, of a basic and a more advanced kind. Function (3) is thought to be an associative process: given the output of a type II module (a word, a picture, a chair) it brings up the relevant information about that kind of thing. For example, the output of a visual (type II) module might be a 3D representation of a chair. The central system would assign a name, bring up its functions, its relations to other objects, etc. via an associative process rather than a strategic one, that is, it is automatic rather than effortful, and mostly unconscious. Function (4), planning, is a strategic process. It must set a goal, decide on a way to accomplish this, se-

lect the appropriate actions, monitor progress, and verify the outcome.[37] This may be accomplished by several central systems acting in concert, so selecting a goal and strategy uses function (3) and organizing action sequences requires function (2). ([Schwartz, 1990], p. 26) Most of the processing involved in all of the central systems discussed is unconscious. What, then, is consciousness?

> Consciousness can be identified with the phenomenal experience of the contents and operation of a limited capacity central system. This system can also control cognitive processes to some extent by selectively allocating attention to some mental representations and cognitive processes at the expense of others. ([Schwartz, 1990], p. 45)

Wherever attention is directed, there goes consciousness. Consciousness is like a central processor that receives and processes the outputs of the many unconscious modules.

Taken together, Fodor, Norman, Shallice, Moscovitch and Umilta have constructed a formidable model of the mind. It seems necessary that some version of the modularity thesis be true, if only from the neuropsychological evidence. That different areas of the brain perform different functions, that some of these functions are qualitatively different, that most of the processing that goes on in the brain is unconscious, these facts seem to point to some form of modular organization within the brain. With just a bit more work, we will be ready to conceptualize just how to get mind out of meat. A few issues remain to be discussed, including information processing, connectionism and computer-based models of mind, and some of the rest of the many useful concepts which fall in the domain of cognitive psychology.

[37] If this sounds familiar, it is because it is very much like both Norman & Shallices' supervisory system and the discussion of problem solving in the "Findings of Neuropsychology" section (3.2.1).

## 3.3 Cognitive Psychology

Cognitive psychology, very basically, is an approach to human cognition as information processing. It is the study of knowing: memory, learning, thinking, judging, perceiving. It involves observing and testing people's cognitive abilities to determine a kind of functional taxonomy and then positing mechanisms which accomplish the information processing necessary to achieve these abilities. Cognitive psychology draws from many disparate fields, including linguistics, computer science, anthropology and philosophy. It is an opportunistic methodology in this sense. Cognitive psychologists are generally concerned with functional descriptions of mental processes rather than physical ones—the interesting questions are answerable by describing the functions performed rather than the machinery which performs them. Cognitive psychologists believe that there are truths which can be captured at some levels which cannot be captured at others. Inference to the best explanation is a frequent tool in describing how things work. What has this functionalist, information processing methodology discovered about human cognition? What progress has it made that the other branches of psychology should acknowledge? Cognitive psychology, by testing the limits of human cognition and perception, has discovered many of the limitations which are (paradoxically) our strengths as information processors. These limitations must be recognized by any model of human cognition and should be used to guide the construction of a model of consciousness. In the next section I will discuss the findings of cognitive psychology, both from experimental cognitive psychology data derived from humans and from the domain of information processing. The necessary conclusions from information science and from human cognition data will lead us to some important conclusions about thought: that the brain must use some kind of parallel distributed processing to accomplish what it does, that humans use heuristics to make decisions, that the brain somehow produces

intelligent behavior from individually unintelligent masses of neurons, that human perception is largely categorical, and that all of our fallible and often seemingly irrational thinking ends up being robustly intelligent—far more so than the fastest computers.

### 3.3.1 Findings of Cognitive Psychology

Cognitive psychologists have studied how we acquire knowledge, how we recall knowledge, and how we use knowledge. Within the domain of acquisition are perception, the transformation of outside inputs into useful categorized information; and attention, the focusing of perception onto distinct parts of the external and internal worlds. Recall of knowledge and the use of knowledge are intimately linked in many ways, as what is recalled or recallable determines what knowledge can be used. Therefore, memory and thought will rely on many of the same kinds of mechanisms. In this section I will demonstrate some of the findings of cognitive psychology within each of these domains.

#### Perception: Feature nets and Categorical perception

Human beings can see and recognize a vast number of objects. We can read, we can drive, we can recognize "Happy Birthday". How do we do this? One hypothesis is that we use our experience and knowledge to guide us in our perceptions of new things: we expect to see certain things in certain contexts, and these expectations help us perceive things, even if they are ambiguous. This makes us perceive things in patterns, and fits new information into these patterns. At the same time, however, these expectations diminish our contact with reality, because they help us to see what we expect to see. How does this process work? What are its implications, and how far does it reach?

One mechanism which can explain how knowledge can guide perception is the

feature-net. This explanation assumes that we recognize things by seeing first their features, like color, shape, or pitch, and then combine the features recognized into a whole thing: a tree, a word or a song. What evidence leads us to believe that we perceive things through features? There are practical considerations: To recognize things as combinations of features, we need a relatively small number of feature detectors, like line detectors, curve detectors, blue detectors. To recognize whole objects, we would need an impossibly large number of detectors - one for each thing we can recognize! That would be quite impossible.

There are also experimental data which point toward a feature-recognition system, like the visual search data: it is easy to pick out a "Z" among round letters like "O" and "Q", and relatively hard to pick out a "Q" among "O"s, "C"s and "D"s, possibly because when you look for something its particular detectors are primed, and so fire more easily and stronger. There is evidence for categorical perception from a voice onset time (VOT) study: subjects listened to presentations of the phonemes (/ba/) and (/pa/), which differ only in the length of the VOT, with varying VOTs created by a computer, and were asked to judge what they heard. As the VOT increased from 0 msec (/ba/) to around 90 msec, subjects said the sound they heard was definitely /ba/, even though the actual sound they heard varied over a wide range. As the VOT increased from 90 msec to 180 msec, subjects said the sound was definitely /pa/. A smooth variation in the actual sounds presented was transformed into a dichotomy, such that a very small physical difference in the VOT made a large difference in the judgment of the phoneme. Thus we have evidence for either a /ba/ vs. /pa/ feature detector or a general VOT feature detector, capable of categorizing VOTs.

Two more kinds of evidence point to the existence of categorical perception: restoration and regularization. These are both evidence of graceful degradation, the gradual loss of function with degraded input. Restoration is the creation of

complete perceptions from incomplete data, so that a presentation of a word with one letter blocked out or incompletely drawn is restored to the complete word. If the presentation is done tachistoscopically, subjects will not notice the incomplete letter. Regularization is the transformation of irregular data into perception of regular data, so that if the letters "CQRN" are presented tachistoscopically, subjects will say that they saw "CORN" even though they did not. Both of these effects are created by categorical perception, which makes us very robust feature detectors with the sacrifice of some of the real data. It allows us to read in poor light, to read degraded stimuli, to hear in noisy conditions and across speakers with very different voices, all things that are quite difficult for current computers to do. All of this points to a feature-recognition system.

The feature net is one way to explain feature recognition, or categorical perception. How is categorical perception accomplished? One explanation is the pandemonium model of bottom-up processing. In this model, low-level detectors all compete with each other to pass on their decisions to the next higher level, with those detectors that are used most frequently having a greater chance than the less frequently used detectors. A metaphor for this process is a group of homunculi, all shouting when their particular feature is detected. Those that detect the most common stimuli shout the loudest, so that only in clear and unusual cases will the less-used homunculi's shouts get through. This model can explain regularization and restoration quite well: when a stimulus is ambiguous or hard to perceive, the loudest (because most often correct) homunculi will win. "CO" bigram detectors will usually win out over "CQ" detectors in uncertain situations, because "CO" is a COmmon letter COmbination whereas "CQ" is not. This makes this system good at dealing with suboptimal stimuli while at the same time diminishing contact with reality, thus explaining regularization and restoration and contributing to graceful degradation rather than complete breakdown with suboptimal stimuli. Here is an

example of reading the word "CAT": The visual system detects a bottom-curve and a top curve and a gap, these features combine to make the "C" detector fire, and "C" is sent to the next level of processing, perhaps still competing with the "O" detector. Simultaneously, other detectors are firing which lead the "A" detector and the "T" detector to fire. The "CA" detector then fires, as does the "AT" detector, and the whole thing is sent somewhere else - you have just read the word "CAT". This system is very robust, that is, it can detect the word "CAT" in a wide variety of situations, even ones where the word is very hard to see. It can do this because it responds to experience, in effect, by "remembering" things it has seen before. How can it "remember"? Two effects, frequency and recency, help explain this.

The feature net's memory is one way that knowledge guides perception. The more frequently or recently a detector has fired, the easier it is to make it fire again. Because you have seen the letter combination "TH" many times before, it is easier to see it again, much easier to see than "TG", which you have almost never seen. This effect is called "priming" - your "TH" detector is more primed than your "TG" detector. So your feature net "knows" that you are more likely to see one than the other. This makes you able to see the word "THE", even if it is showed for only 20 msec, or if part of a letter is missing, because the relevant detectors are so highly primed that only a weak stimulus will cause them to fire.[38] Priming, however, also leads you to make mistakes, so that if you are presented with "CQAT" for a short period of time, and asked what word you saw, you will say "COAT", because the weak stimulus was enough to make the "CO" detector fire and not enough to make the "CQ" detector fire. Your knowledge has guided your perception, but you have been lead astray. In this way, the perceptual system helps you to see things in ambiguous situations while diminishing your contact with reality. These effects are

[38]This is, for once in cognitive psychology, quite neurophysiologically plausible: permanent alteration of synaptic action and long-term potentiation of synaptic action are quite possible.

not confined to visual perception of words: Your auditory system also appears to use a feature net. If presented with an ambiguous input, such as someone saying the word "legislation" but with a burst of noise coinciding with the "s", you are able to hear the word. Also, you don't know what happened: you know there was a noise, but you don't know when it occurred or which letter it blocked out.

How far does this feature net go? How much does your knowledge guide your perceptions and diminish your contact with reality? There is evidence of bi-gram (two-letter) detectors, but could there be word detectors? Word combination detectors? The answer to the last two questions has to be no, for purely practical reasons: a feature net with bi-gram detectors needs 676 (26 x 26) detectors, a manageable number. But to have word detectors, or word combination detectors, the numbers become astronomical. Nevertheless, you can recognize words in context better than by themselves, and other methods of priming work to make you able to see ambiguous stimuli. If told to think of a round, red thing, you are more able to see the word "apple." These two examples can't be handled by our feature net: they seem to be examples of top-down processing, of complex knowledge of categories influencing what we see.

Clearly, the feature net does a good job of explaining certain parts of perception, of pattern recognition, of our ability to read. However, it doesn't explain everything. The fact that we can recognize syntax and that context and concepts can also influence perception all call for a complex, top-down explanation, one that includes the abilities to search for information in different categories and use that information to prime the relevant detectors in the perceptual systems. All of this also has a fairly frightening implication: In effect, we see what we are used to seeing and what we want to see!

**Attention: Divisible resources and Automaticity**

Attention is another function cognitive psychology has attempted to explain. One approach to understanding attention is finding its limits, testing its weaknesses, seeing how attentional mechanisms perform when they are forced to work on multiple tasks. To do this, a divided attention task can be used, like dichotic listening. In this task, two separate stimuli are presented, one to each ear. How much of the information can be assimilated? How much can be recalled? How many things can you pay attention to at the same time? Older ideas of attention included some assumptions about mental resources. The argument goes like this: (1)Mental processes require work. (2)Perceiving is a mental process. (3)Therefore, perceiving requires work. (4)Work requires resources. (5)Resources are limited. (6)Therefore, only a limited amount of work can be accomplished. According to this view, one could divide up attentional resources, giving some percentage to one task and the remainder to another.

However, this view is not quite correct. Some tasks are more difficult, some are easier, and some, though they seem quite difficult (and are difficult to learn) become automatic with practice. Therefore, two other kinds of resource theory have been proposed: specific resource theory, which says that the amounts of resources available depend on what specific task is required, and channel segregation theory, which explains attention division in terms of separate channels of information processing. If the required tasks are quite different, or call on very different types of resources, then attention can be more easily divided. If the tasks call on the same resource or kind of resource, then the general resource theory is a good explanation. If the task needs to have separate channels of attention while avoiding crosstalk between them, then channel segregation theory is a good explanation. In short, what matters in divided attention—what limits it—is whether the tasks involved call on the same resources.

There is one more effect to be explained within this area: automaticity, or the effects of practice. A good example of automaticity is driving a car: it requires paying attention to many things simultaneously, like the road, other cars, speed, traffic lights, pedestrians, etc. Most people can do all of this and hold a conversation at the same time, thus dividing their attention between many motor, spatial attention, and computational tasks (all for driving) while at the same time comprehending speech and replying to it intelligently. How can this be accounted for? Automatic performance seems to require no attention at all, or very little. When the task doesn't vary greatly from time to time, it can become memorized so well that no "thought resources" are required to perform it.

The most recent work on attention has been on the "response selector" mechanism of Harold Pashler [Pashler, 1993]. This mechanism is proposed to act to initiate all responses, so it is a kind of general resource and therefore a limiting factor in attention. This theory places the bottleneck in information processing not in perception nor in an inability to do two motor responses simultaneously (like speaking and pushing a button) but with the existence of a single mechanism that is responsible for choosing responses. Like most cognitive experiments, subjects are given a very difficult task which tests their limits. In this case, the task involves two separate stimulus-response pairs, in which the two stimuli are presented and the subject must respond as quickly as possible to each. Response time for the first stimulus is usually a constant, but response time for the second task is much longer. Why does the second response take so long? There are three possibilities: the perception of the second stimulus, the production of the response, or the selection of the response could be a bottleneck. By increasing the difficulty of these three factors it can be determined which stage is creating the bottleneck. Pashler's experiments have shown that increasing the difficulty of the perceptual or production stages (by making the second stimulus hard to see, or the response of the first task

more difficult) has no effect on response time for the second response, but making the selection of the response more difficult added a constant to the reaction time to the second stimulus. In other words, there seems to be some sort of response selector mechanism which can only initiate one response at a time, and is slowed down by making responses more difficult.

Thus far we have explored constraints on perception and attention. Perception was seen to be categorical in nature, which confers advantages in perceiving degraded stimuli at the expense of contact with reality. Attention constraints depend on the tasks at hand: if they use the same resources, then they must compete for that resource as a solitary amount. With tasks that require different resources, separate task specific resource limitations define the constraints. With automaticity, however, the resources seem not to be used at all—automatic behaviors require little or no mental effort. Finally, in all cases response selection appears to constrain how many things can be performed at once. In the next sections we will examine short- and long-term memory and some of the heuristics people use in decision making.

### Memory: Short-term and Long-term

What is short-term memory (now called working memory)? What does it have to do with speech? These two questions will be answered along with an account of how they might interact. The old picture of short-term memory will be re-evaluated, in light of some new data which show that the old account was too simple because it didn't include some important ways that memories are held in working memory, and some data which show that speech has an important effect on how working memory works.

The old picture of working memory was fairly simple: information came in through the senses, was categorized, and sent to a 'loading dock,' or information was retrieved from permanent storage and put on the dock. Once on the loading

dock, these bits of information either were moved into permanent storage, were held on the dock, or were pushed off the dock into the garbage, never to be seen again. The 'loading dock' is the working memory, where all work is done on information, whether that be arithmetic, philosophy, or simple rehearsal. For a time, this was the picture of working memory: it was a good theory, and it seemed to fit the data. However, it ran into a few problems, and a more complex system was postulated, one with a central executive and two helpers, each of which helped maintain things in memory. One helper would maintain visual images, while the other worked on maintaining verbal information.

Baddeley and Wilson [Baddeley and Wilson, 1985] did a study which bears on the verbal helper, and they came up with an idea of what the rehearsal system was doing and how it does it, called the phonological loop. On page 181 of their paper, they explain their hypothesis like this: "The Phonological Loop is assumed to comprise two components, namely a temporary phonological store linked in with an articulatory rehearsal process." In other words, there is a 'remembering' part and part which does speech planning and rehearsal, which is connected to vocalization. This system posits that overt or covert subvocalization is actually part of the rehearsal process, so that speech production becomes an integral part of the working memory picture. It also raises some questions: is it harder to use working memory while talking, or even just making any kind of noise?

Working memory can hold about seven bits of information at a time. If rehearsal through speech is part of the way working memory holds on to information, than any sort of vocalization might interfere with this process, perhaps reducing the holding capacity of working memory. If this was the case, it would give us reason to believe that Baddeley is right, that the speech process is included in working memory. There are several different effects which tell us that perhaps Baddeley is on the right track: First, the phonological similarity effect, which shows that it is harder

to remember a sequence of similar sounding words, like "cat bat hat flat map tan," than a sequence of dissimilar words, like "dog pill joke cow day." Words that sound the same get messed up in the speech process: try speaking the words in the first example out loud (or to yourself) a few times to get a feel for the effect. Second, there is the irrelevant speech effect, which shows that the presentation of spoken material, even in another language, interferes with the ability to remember visually presented items, like strings of digits. Most impressive, I think, is the articulatory suppression effect. When the subject has to say something over and over, like "the," it interferes greatly with the ability to hold items in working memory, from about 7 to about 3. Furthermore, with concurrent articulation the phonological similarity effect and the irrelevant speech effect disappear, adding more reasons to believe in the phonological loop model of working memory.

Clearly, however, people are not talking to themselves while using working memory. Does this discount the phonological loop model? Baddeley and Wilson [Baddeley and Wilson, 1985] tested a subject who could not speak because of brain damage, but still had language skills. He showed the phonological similarity and word length effects, so clearly actual vocalization is not needed for rehearsal. The explanation for this is that the phonological loop employs a fairly deep-level speech generation system, rather than the actual muscles involved in the speech process. Baddeley has given us a convincing new picture of working memory's rehearsal process. Both the experimental data and the explanatory success of his theory indicate that it is true, that articulatory processes are implicated in working memory.

Now we move on to long-term memory. How long does it last? What are the constraints on memory recall, especially after a long period? Here are summaries of four studies on memory retention, which will demonstrate some of the parameters of long term memory retention.

Bahrick et. al. did two seminal studies in this field, and he was the first person

to give the name permastore to long-term memory, based on the results of his experiments. Bahrick's first study [Bahrick et al., 1975] was a test of people's memories for their high-school classmates names and faces, tested at several retention intervals (RIs). He did two different kinds of tests, free recall and recognition (which we would call explicit and implicit). He found that, for the free recall tests, people's memories showed a steady decline, with an accelerated final decline of memory in the subjects' old age. However, tests of name-face matching and name and face recognition showed that people's memories remained at 80% for about fifteen years, then started to decline, and on tests of face recognition people got 90% right for about 35 years before a final decline.

Bahrick's second study [Bahrick, 1984] was of people's memories for Spanish learned in high school. The results could also be correlated with how well they had learned the material in the first place, because their grades were available. The results of this study were very impressive, with the best subjects getting 80% correct on recognition tests after 50 years. The subjects who had gotten good grades in high school did much better than subjects who hadn't, but their memories also remained stable up to the fifty year RI. These studies are what prompted Bahrick to believe there was some kind of "permastore." He postulated that memories were permanent until the effects of old age took their toll.

Neisser [Neisser, 1984], however, didn't believe that Bahrick's data necessarily implied any kind of "permastore." He argued that, instead of specific knowledge being stored permanently, knowledge schemata were created. These schemata could create the knowledge needed, when needed, and could account for the data in a different way. This hypothesis makes certain predictions, one of which is that memories for concepts will be remembered a long time, because they are schema-central, but specific knowledge will be forgotten.

Conway, Cohen and Stanhope ([Conway et al., 1991], hereafter CCS) tried to

replicate Bahrick's study, and at the same time test Neisser's hypothesis. They did a study on students' retention of cognitive psychology learned in college over various RIs, and they also did implicit and explicit memory tests. Furthermore, they knew exactly what the students had learned, so they could test for memories of specific details and of general concepts. The results of this study also support the permastore hypothesis: even after very long RIs (125 months), subjects' performance on memory tests was well above chance. They noticed an interesting pattern in the subjects' memories - their memory would decline until 36 months, after which they would remain stable up to 125 months. The CCS study also disconfirms Neisser's hypothesis: comparing results of recognition tests for names (specific information) and concepts (schematic information) we see three things. First, retention was about the same for both kinds of knowledge, although names did drop off faster; second, retention for both was much better than chance; and finally, retention was stable for both until an RI of 125 months. The authors say, at the end of their paper, "We conclude that, in the present study, knowledge was retrieved from memory, rather than reconstructed by schemata."

So far, we have seen a line of evidence that points toward something like permastore as a model for memory. Even over RIs of fifty years, information is reliably retained. The CCS study showed that even 'peripheral' knowledge is retained, knowledge for proper names or dates. Bahrick's studies showed that people have good memory for names and faces and for Spanish, even after 50 years. It is actual knowledge that is retained, not just schemata. Does this answer all our questions about memory? Of course not, nor does it answer all our questions about permastore.

The main problem with these experiments is that the really impressive results were from implicit memory tests, not explicit tests. In other words, if you asked subjects to recall some information, like saying "Please give me a glass of water" in

Spanish, they probably wouldn't be able to do it after a sufficiently long RI. The information seems to be in their heads, they just can't consciously access it, and as we all know, unavailable knowledge doesn't do much good at all, except on implicit memory tests. It certainly seems like there is some kind of memory "permastore," but it is not necessarily useful, and feels like the information is gone. How accurate, over the long term, is ordinary remembering? Well, according to these data, it is very accurate when tested in the right way, and not so accurate when tested explicitly. In the next section, the importance of availability for recall will be examined as a heuristic used in decision making, along with representativeness.

### Decision making: Human irrationality

Humans are not optimal decision makers. This may come as no surprise, except that most people are quite confident that the decisions they make are entirely rational. Perfect rationality, however, has tremendous costs associated with it: it can be incredibly slow and difficult, and therefore not adaptive. Time and energy are important constraints on all biological functions, and appear to be so for human information processing as well. While people seem to do very well, however, there are many domains in which they can be shown to be irrational, overconfident or overhasty. These apparent logical failures are a large part of human thought, and one of the more important contributions cognitive psychology has made to understanding how people think.

Rationality and logic provide prescriptive normative standards for decision making. Analysis of possible outcomes should be made according to statistics. Self-contradiction should not occur. Judgments should be made not just on what first springs to mind, but on a careful weighing of all available information. Probability should not be ignored in favor of similarity to a schema. While all these prescriptive standards seem reasonable, humans have been shown to fail to take any of them

into account in making decisions.

Tversky and Kahneman [Tversky and Kahneman, 1981] did a study asking people to make a choice about a hypothetical epidemic. The choices were an 80% probability of losing 100 lives vs. a sure loss of 75 lives. They preferred the first choice rather than the certain loss. When asked to choose between a 10% chance of losing 75 lives and an 8% chance of losing 100 lives, they chose the 10% chance of losing 75 lives. These two sets of choices are the same, basically, although the first set involves certainty.

Tversky and Kahneman [Tversky and Kahneman, 1974] did another study which demonstrated not just ignorance of statistics but self-contradiction. This is the famous framing effect study, in which the way the question is posed changes people's responses. The question asks:

> 1. Imagine that the U.S. is preparing for the outbreak of a disease which is expected to kill 600 people. Two alternative programs have been proposed. Assume that the exact scientific estimate of the consequences of the programs is as follows:
>
>    - If program A is adopted, 200 people will be saved
>    - If program B is adopted, there is a 1/3 probability that 600 people will be saved and a 2/3 probability that no people will be saved.
>
> Which of the two programs would you favor?

An alternative version of the problem is given to another group of subjects, for whom the situation is the same, but the choices are:

- If program C is adopted, 400 people will die.
- If program D is adopted, there is a 1/3 probability that nobody will die, and a 2/3 probability that 600 people will die.

Subjects reliably choose A over B and D over C. This is not logical, as the two sets of questions present the exact same choices, just phrased differently. The way the question is framed affects the way people answer it—people over-value losses and over-value certainty, contrary to the prescriptive, normative rules of probability.

The availability heuristic describes another failure of humans to live up to the prescriptive rules. This heuristic is just choosing what easily comes to mind (what is available) rather than thoroughly thinking through decisions. When asked to decide whether there are more words in the English language which (1) start with the letter "R" or (2) have the letter "R" in the third position, most people will say (1). This is possibly because it is much easier to come up with a large number of words that start with "R" than to think of all words with "R" in the third position, even though it is not the case.[39]

I will only discuss one more common normative failure, although there are many more. This common judgment strategy is called the representativeness heuristic. It is the judgment of probability by matching to schemas. In the conjunction fallacy, people will often choose as most probable a group of qualities which match a stereotypical picture of a typical case. For example, when asked which of the following is most likely:

1. That a man is under 55 and has a heart attack
2. That a man has a heart attack
3. That a man smokes and has a heart attack
4. That a man is over 55 and has a heart attack [Medin and Ross, 1992]

people will often choose (3) or (4), because those cases are closest to the representation of a heart attack victim in most people's minds. Of course, (2) is the most

[39]If you don't believe this, take a quick look through any paragraph in this text (for a sample of regularly used words, although it's not particularly representative), or search through a dictionary.

likely, because the other choices involve added probabilities beyond maleness, and so are actually less probable.

Clearly, people do not use normative standards in making decisions. Many choices are made from heuristics which work fairly often but are not logically normative, although they are probably adaptively normative because they are more efficient—over a short period of time, heuristics are much more successful than normatively correct procedures. These facts about human judgment must be taken into account by any models of human information processing, and should inform us about some of the ways the mind works in general.

### Computer Metaphors & Serial vs. Parallel Information Processing

Notions of information processing have changed somewhat since the early days of computer science. The paradigm for any sort of computation used to be the Turing universal computer, a device of any physical description[40] which takes in information one piece at a time, finds a rule based on what state it is in and carries out that rule. For example, it might be a device which reads a paper tape covered with ones and zeroes, and have a few very simple rules like "if the tape says 1, move three digits to the left; if the tape says 0, erase that digit, write a 1, and move two spaces left." With this kind of organization, enormously complicated systems can be constructed, which seem quite sophisticated, like personal computers and compact disk players. This kind of computational device has been used to model the mind, using a central processing unit (CPU), a machine table (which contains the "rules"), and a flexible memory space. There are, however, many good reasons to believe that the brain does not work this way.

First of all, this process is rigid. It doesn't make mistakes, it doesn't guess,

[40]That is, its essence is contained in its functional description, which can be instantiated by any physical system.

it methodically carries out its instructions based on its inputs and its machine table states. Second, this process is serial. The CPU can only do one thing at a time, whereas humans seem to be able to understand many things at once, within limits. Third, even though computers are able to perform calculations far faster than humans, they cannot match our flexibility in information processing. The disparity in speed of processing is truly enormous: in the time it takes for just a few synapses to fire, say enough to react reflexively to a stimulus, a computer can perform thousands of calculations. This speed disparity is one of the most important reasons for believing that the brain is a parallel processor rather than a serial one: in order to process the requisite about of information for a task many calculations must occur, but the limitations of synaptic transmission mean that there is only time for a few bursts of synaptic activity. Since these calculations could not be accomplished by those few bursts of activity working serially, one at a time, the brain must be doing something else. Fourth, computer models of human cognition based on the way that computers work run into what is known as the knowledge problem. The knowledge problem involves how one accesses information, particularly the relevant information.

One approach tried by early Artificial Intelligence (AI) researchers was to explicitly place the requisite knowledge into the model. By this method they created expert systems, computer models that were good at only one kind of task. The hope was that by creating several of these systems and hooking them up, a model of intelligent behavior could be created. Unfortunately this was not the case: these types of models all run into the knowledge problem—they cannot decide what information is important for any particular problem. Humans seem to have content-addressable memory, whereas computers must have explicit lists of memory locations unorganized by semantic information. An example may help to illustrate this. We have no trouble interpreting this simple story:

> Betsy wanted to bring Jacob a present. She shook her piggy bank. It made no sound. She went to look for her mother. ([Schwartz and Reisberg, 1991], p. 330)

because we know that presents cost money, that money is kept in piggy banks, that the money in piggy banks is generally made of metal and so makes noise when the bank is shaken, that to determine the contents of a piggy bank it must be opened or shaken. All of this knowledge is seamlessly applied to the problem of understanding that simple story, knowledge from many domains. How do we decide what is relevant?

These considerations have led many people to believe that the brain processes information in parallel, rather than serially, and that knowledge is stored in a very different way in the brain than it is stored in a computer. Parallel processing allows many computations to happen simultaneously, in different processors, and then lead to an output much faster. Taking neurons for a model, some AI researchers have created neural network models of information processing. These networks are composed of (virtual) neuron-like units, called nodes, each of which is connected to many others. They process information by summing the inputs of other nodes, and either reaching an activation threshold and then sending off impulses to those nodes to which they are connected or doing nothing. Like neurons, these inputs and outputs are multivalent: they can be negative or positive, strong or weak, on a continuum. In these models, knowledge is stored in the connection strengths between nodes, rather than in a separate storage place. No single node stores information, rather the information is distributed across the entire network. This kind of distributed parallel processing model is indicated not just by neuroscience and neuropsychology but also by the constraints of information science, by facts about computation and data storage, and so could help us to understand how the brain can think.

### 3.3.2 Methods of Cognitive Psychology

The methods used by computer scientists in modeling information processing are part and parcel with their findings: whether a model works and how it is constructed are both based entirely on success and computability. The data from humans, however, requires some explanation. How have cognitive psychologists gathered these data? Is it trustworthy? I will discuss a few of the methods used by cognitive psychologists in the next section, including reaction time (RT) measures, implicit and explicit memory probing and priming, and the artificial constraints of the laboratory. In addition, I will provide a short discussion of the method of inference to the best explanation and the reason cognitive psychologists depend on it. Finally, two pictures of human information processing from cognitive psychology and computer modeling will be examined, before we move on to Dennett's grand synthesis and my own.

Reaction time measures are important in cognitive psychology tasks for several reasons. First, they provide a good, easily analyzed quantitative measure. Second, they are relatively easy to acquire and control, especially in tasks where computer presentation of stimuli is possible. Finally, and most importantly, they get at the limitations of human performance. Finding out how quickly a task can be performed gives us clues to how hard the task is, and therefore how much information processing must occur to produce a response. Reaction times and response latencies can be used in perception tasks, attention tasks, and memory tasks, by designing experiments which test the constraints imposed by human thinking on each of these functions.

Implicit and explicit probes can be useful in discovering how information is processed and retrieved. These kinds of measures can be used in juxtaposition for the same task, as in memory retrieval. For example, when shown some Turkish script characters and asked which ones were previously presented, memory often fails and performance is at chance. However, when asked which characters are preferred, some

other method of retrieval is used and performance improves dramatically, because people seem to prefer things they have seen before. Implicit stimuli (or subliminal stimuli) can be a useful tool for determining how much of the information we receive is processed unconsciously, and so how much of the world around us influences us without our knowing it. These kinds of tests, especially when used in contrast on the same tasks, can be used to examine top-down influence on perception, expectation effects on attention, and memory retrieval. With them we can learn how perception is biased by previous information, how attention is focused by implicitly perceived stimuli, how information can be accessed by different search techniques.

Imposing artificial demands for some tasks is another way of probing the capacities and functions of the mind. By making tasks very basic and very difficult, specific functions may be tested to their extremes. In this way cognitive psychologists can examine the parameters of cognition and break it down into its component parts. Once some law-like relation is found, like categorical perception in word perception, that relation may be applied to other cognitive functions to test it and determine its generality. Finding out where people make errors is another important method for cognitive psychology. These errors come out more clearly with difficult tasks meant to challenge subjects functioning. Through an analysis of the kinds of errors and limitations inherent in human information processing a picture of human cognition can emerge.

Finally, cognitive psychologists must explain their data. What methods do they use? One of the most important kinds of explanation in cognitive psychology[41] is inference to the best explanation. This means creating an explanation based on the inputs and outputs of something, and then inferring how it works. If many of the functions of a particular kind of cognition can be discovered, then a functional description can be created. This kind of description may be a model, which may

[41] And in other sciences, whether or not they'd care to admit it

make predictions about other untested functions. For example, perception of words takes written text as input, and produces semantic understanding. After much testing at the input level, a model of categorical perception was created, with feature detectors and pandemonium as a functional description. This model then can be applied to the other senses, and the predictions it makes about their functions can be tested. This can be a very powerful method for understanding a system as complex as the brain, especially if constraints from information processing and neuroscience and neuropsychology are used to guide explanation.

### 3.3.3 Information Processing Models of mind: Flowcharts and PDP

One of the ways cognitive psychologists have modeled the mind is by segregating its functions and then integrating them into a complete picture. These pictures are called flowcharts, and present the mind as a series of mechanisms whose functions have been mapped out by cognitive experiments but whose neural structure is unknown. 3.5 shows a simple version of one of these models, including the feature detectors of perception (called filters in the model), a very short-term storage place (in vision, iconic memory; in audition, echoic memory; in the model simply "buffer"), working memory, the phonological loop and its connections to speech, and long-term memory storage.

These kinds of models can be productive in many ways, especially by stimulating new questions about each of the processes. Once a group of cognitive psychologists agree on a certain functional part, like the rehearsal loop, more specific questions can be asked. Is there a similar loop for visual processes? How does this loop affect working memory? In addition, these models can provide questions for neuroscientists and neuropsychologists: How could these processes be physically instantiated in the brain? What neural architecture could perform the function of a "buffer"? Func-

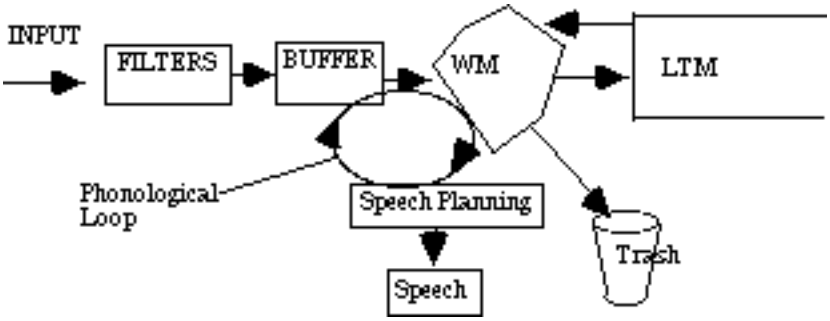


Figure 3.5: **A Simple Information Processing Flowchart**

tional characterizations of sensory processing can aid in understanding how certain perceptual effects occur, and point toward the kinds of neural circuitry that could accomplish these information processing tasks. These models also do a good job of explaining some pathologies: anterograde amnesia, for example, can be explained by a broken connection between long-term and working memory—information stored in long-term memory can be loaded into working memory, but new information can no longer flow into long-term memory. Disorders at the filter or buffer levels can explain some neuropsychological deficits.

These flowcharts, and the "boxological" style of theorizing, have some large explanatory gaps, especially in explaining consciousness. How does the "working memory" work? Where (or what, or who) is the central executive, and how does it work? Cognitive psychology tells us some of the functions such a mechanism must perform, but leaves out much of what seems important. Theories which contain a central executive are especially disappointing, because what is the central executive but an extra intelligence in the head, a homunculus, a Cartesian viewer? How does his head work? The tendency to get much of the perceptual processing story right but cling to the Cartesian viewpoint for the interpretation of this data is unfortunate. Why should all the work these higher-level perceptual processes have accomplished be reinterpreted by a central executive? These are the ideas Daniel Dennett fights

against quite vigorously in Consciousness Explained [Dennett, 1991], which we will examine in the next chapter after a discussion of the latest attempts at computer models of the mind from cognitive and information science.

The model from computer science I will present is based on the ideas of parallel processing and distributed knowledge, as well as including aspects of pandemonium. This model is called Parallel Distributed Processing, or PDP. It is a line of research which has grown out of neuroscience and computer science, and its aim is to model the brain. It does this by using extremely simple components, analogous to neurons, called nodes, and interconnecting them to a high degree. It is called parallel processing because at any one time many of the nodes are firing at once, and it is distributed processing because instead of the nodes holding any information, it is the connections between them that hold the information. These models are capable of learning by the alteration of the connection weights between them. This is done by back-propagation of error messages, in a process called supervised learning.

The strengths of this style of theorizing are several, and most of them are very persuasive. First of all, PDP theorizing must be right in some respects because we know that it is similar to the actual biology of the brain. Neurons are nothing more than "spit-catchers and spitters," they don't hold any knowledge in themselves.[42] PDP captures this truth. Furthermore, we have strong evidence for the theory that the brain works in parallel: we know how long it takes for neurons to fire, and we know how much time it takes to have a thought or make a decision. Combining these facts, we arrive at the conclusion that there is only time for twelve to fifteen steps of processing. In order for our thoughts to have enough content to be smart, there must be lots of them firing at each step, therefore, the brain must be a parallel processor. The brain probably also stores its knowledge in a distributed fashion, perhaps even in the connections between neurons, and PDP incorporates this fact

[42]That is, there is no single neuron where the word "apple" is stored.

as well.

Another strength of PDP theorizing is that we can see every step of the process. With the human brain, thinking is, for all important respects, an invisible process. With PDP models, we can see every step of a 'thought' process because they are done on a computer. We can collect all that data, and tell what every component of the system is doing and know the connection weights between every node. We can't perform such analysis on a human brain.

The last strength of the PDP style of theorizing is the fact that it has been successful in several respects. The fact that a PDP net is capable of learning things and applying that knowledge is amazing - starting from random connections, a net can fairly quickly learn to add or subtract. Two examples of practical models that exist are a net that analyzes sonar input and can detect submarines, and a net that can read aloud written English. All of the above are actual strengths of PDP style theorizing. It also has a potential strength, that with a good enough model of the human brain it will be able to explain exactly how the brain works, much more completely and accurately than psychology has to date.

Unfortunately, PDP has some problems. It claims that, with a completed theory of neuroscience, psychology will become a bankrupt enterprise, which is working at a level of complexity which is, by the way, arbitrary. This position has caused, not surprisingly, a considerable backlash among psychologists. PDP has been shown to run into the same practical problems that expert systems ran into in the early 70's. An expert system is a heuristically guided program, good at one very limited thing, like medical diagnosis. These systems avoided the knowledge problem (the fact that once a computer has enough information stored to understand simple stories it bogs down in memory retrieval) by severely limiting their focus, i.e. to drug prescription for various medical diagnoses. It was thought that connecting several of these expert systems linked together would solve the knowledge problem, but it didn't - several

expert systems linked together have such a huge database that once again they bog down in retrieval. PDP has the same problems: so far, any one net can only do one thing, like adding 1 to a number but not being able to add 2, and no-one has figured out how to make anything other than these "toy" programs.

Another problem with PDP is more troubling because it has to do with the way that it explains thought. PDP has a good thing going for it: it can tell you what every single 'neuron' did in a given thought process, and describe it as a point in n-dimensional 'weight space,' determined by the positions of every single connection weight within the net, or it can give you a list of these connection weights. The problem is, this is no explanation. "Why did this net successfully add 2+2?" "Here, look at this list of 40,000,000 connections." PDP doesn't tell you why it's that particular list, or if there are other lists that would do the same job.[43]

Potentially, PDP can model the human brain. We know it is fundamentally right in some respects, with regards to the biology. But it also fails in certain ways: it must be able to have good explanations, not lists; it has to explain how neurons exhibit rule-governed behaviors; and it has to be able to have a multi-purpose net, rather than just limited, 'toy' nets. Even if PDP succeeds in solving all these problems, however, it may turn out to get rid of psychology in the same way that physics gets rid of biology; that is, not at all. If PDP theorizing is used in interesting ways, it can help us understand how the brain works, but it will probably never replace psychology at certain levels of analysis.

## 3.4 The Incredible Machine

From neuroscience, cognitive neuropsychology, cognitive psychology and cognitive science we have learned a great deal about the workings of the brain, both its

[43]This is much like the reasons I argued that explanations at the level of neuroscience are not good explanations—and why some functional characterization would be more satisfying.

limitations and its strengths. Neurons, the basic constituents of the brain, are both simple and complex, and exceedingly plastic. At base they are digital: they either fire or do not fire. The mechanisms of activation and inhibition, however, are exceedingly complex, because of the variety of neurotransmitters and the numbers of synapses on each neuron. Even within the interaction of just two neurons there can be activations and inhibitions of various durations and strengths and permanent or semi-permanent alterations of functioning. When combined into circuits they can accomplish virtually any kind of computation. With recursive loops, timing chains, etc. they can accomplish Fourier transforms of wave functions, vector analysis, color detection under varying lighting conditions—all quickly and unconsciously.

Larger groups of neurons are combined to perform more demanding tasks like word, speech, object, odor, sound, and texture recognition, and all this with the same unconscious speed and accuracy. Many separate groups like these, called modules, can be coordinated by higher level processing modules which integrate their outputs into sophisticated representations of the world and allow humans great flexibility in action and goal-fulfillment. From the information stored in these large networks of neurons the future can be predicted to a surprising extent, such that humans can predict the behaviors of other complex systems in the world. Pattern recognition in both the forms and behaviors of things in the world is performed constantly by multiple systems simultaneously, allowing instant access to complex representation of whatever factors are important to the organism at any time. All of these powers (and the limitations that go with them), it seems to me, have been well explained by psychology at each of the levels I have discussed. Where there are gaps, the necessary information can be guessed at or is being sought out.[44]

What, then, is still missing? Where is the mystery still to be explained, what

[44]I may be called optimistic or naïve, but it seems to me, at least, that the questions that remain are empirically discoverable by the methods we have or are developing (i.e. high-resolution MRI).

are the crucial missing pieces? The answers to these questions require, I believe, a conceptual revolution in the way we think about consciousness, a revolution forced upon us by the things we have discovered about the workings of the mind. Philosophers will haggle about the intrinsically unknowable subjectivity of the mental, but I believe this is a non-problem as far as understanding how we "get mind out of meat." In the next chapter I will discuss Daniel Dennett's answers to these questions and his strategy for overcoming the conceptual difficulties which still block our understanding. In the end, with the mental tools provided by this chapter and the next, we will be able to see how an understanding of consciousness is possible, even if we have not yet solved all of the important philosophical and empirical problems.

# Chapter 4

# Consciousness Explained?

Daniel Dennett, in *Consciousness Explained* [Dennett, 1991], sets out to dismantle what he sees as the primary conceptual stumbling blocks to an understanding of how the brain could produce consciousness. The most pernicious of these is the Cartesian Theater: the idea that there is a place in the brain where all the elements of sensory processing are presented, as if on a screen, for conscious viewing. The limited and serial nature of conscious experience, of episodic memory, seems to indicate that there is a place where "it all comes together," where all of the evidence of our senses is projected simultaneously for conscious viewing. The problem with this view is that it requires reprocessing: after the sensory mechanisms have already processed the raw data of the world into final form, something in the brain then watches the results, a separate intelligence, the Cartesian self. Who is this self? How does it work? The answer is that there is no little man in the brain—by the time information from the world has been processed, all of the elements of conscious experience have already been built into the perceptions by the very systems which processed the information in the first place. Indeed, that building up of meaningless information into meaningful, categorized perception is what those processing systems do. Moreover, there are many of these systems, all operating in concert (in parallel) and it is this multiple simultaneous processing that creates conscious

experience. Dennett's name for this is the Multiple Drafts model of consciousness, based on analogy with the way articles and books are created now. At any one time there are many different drafts of a text in existence, none of which is the canonical version. The final, printed edition is a static, dead thing, mostly useful as an archive.

Dennett's version of mind is bipartite. He believes that the brain is a large parallel processor at the neural level, but that many of its functions are governed by a serial "virtual machine" overlaid on the neural achitecture, a "Joycean machine" which uses and is influenced by words and stories and information from culture. He also believes that consciousness is too recent a development to be evolutionarily explainable, so the explanation must come from a newer, faster process: cultural evolution.

Finally, Dennett explores some of the philosophical implications of his beliefs. He covers mental representation and the reasons it doesn't require an inside observer 'seeing' reconstructions of the external world, more argument against the Cartesian Theater. He explains how mental phenomena can have content without language, which can then give rise to speech acts if they are important content-fixations. In this section of the book he also tries to fight against the persistent intuitions about qualia and zombies[1] and for the positive accounts of perception and consciousness that come from neuroscience and from the Multiple Drafts theory.

In this chapter I will go over Dennett's main points in some detail, especially his points in the more psychological and scientific part of the book. We'll follow him on his exploration of consciousness "from the inside out" ([Mandler, 1993], p. 335)—from the phenomenological, introspective "feels" of consciousness to some methods for investigating these to some psychological, psychophysical, anthropo-

[1]Zombies are philosophical creatures which seem to be human in every way but lack consciousness, have no phenomenal experience, experience no qualia.

logical, biological and computational evidence for believing Dennett's conclusions about consciousness.[2] After getting Dennett's story straight I will mention some of its problems, and then finally present my own conclusions about what consciousness is and how we can explain it.

## 4.1 Phenomenology and Heterophenomenology

What is phenomenology? Why is it important to an understanding of consciousness? Phenomenology is the study of mental phenomena, an attempt to get them all out in sight so that they can be explained. Success in this area is vital to an explanation of consciousness, because mental phenomena are the things which constitute consciousness, and any explanation of our mental life which doesn't include satisfying explanations of the phenomena we experience will not be a satisfying explanation at all. Dennett uses zoology as an analogy for phenomenology ([Dennett, 1991], ch. 3) because it is the same kind of classificatory scheme, and because it points toward a way of reinterpreting the objects we wish to classify.

If I were to say something ridiculous like "there are no animals" I would be called mad. But if I then said "what I really mean is that these things do indeed exist, but they are not *animals* but really just robots, biological mechanisms which we call animals but which should be reinterpreted as just *mechanisms*." Is this so mad? Is this not something like what a zoologist might tell us when asked to explain what animals are? This same kind of classification and reinterpretation of the objects which inhabit our "phenomenological garden" ([Dennett, 1991], p. 45) is what Dennett is attempting.

Dennett's taxonomy of the mental phenomena, his "phenom" (by analogy with the zoologist-zoo metaphor), is divided into three parts:

[2]Throughout this section I will be roughly paraphrasing Dennett.

> (1) *experiences of the "external" world,* such as sights, sounds, smells, slippery and scratchy feelings, feelings of heat and cold, and of the positions of our limbs; (2) *experiences of the purely "internal" world,* such as fantasy images, the inner sights and sounds of daydreaming and talking to yourself, recollections, bright ideas, and sudden hunches; and (3) *experiences of emotion or "affect"* (to use the awkward term favored by psychologists), ranging from bodily pains, tickles, and "sensations" of hunger and thirst, through intermediate emotional storms of anger, joy, hatred, embarrassment, lust, astonishment, to the last corporeal visitations of pride, anxiety, regret, ironic detachment, rue, awe, icy calm. ([Dennett, 1991], p. 45)

All of these phenomena are the objects of conscious experience. The challenge is to explain them, explain them fully, and without reference to any magical properties or substances. Can this be done? Will our explanations of external, internal and affective phenomena leave something important out? This question is about qualia, rather than phenomenology, and so must be postponed. First let us explore the phenomena in greater detail and see exactly what needs to be explained.

Taste and smell are our crudest senses, in that our discriminations of them are less fine-grained than with hearing and vision. They are both triggered by chemicals which fall on the transducers of the tongue and the nasal epithelium. We have very poor spatial and temporal resolution in these modalities, so that odors seem to inhabit large areas and we cannot tell when they were produced. Our sense of taste is very limited, to only four flavors (salty, bitter, sweet, sour) but is linked phenomenally to our sense of smell so that "taste" usually means the combination of flavor and odor, allowing us to distinguish a much larger range of tastes. Similarly, our senses of touch and kinesthesia[3] are phenomenally linked. The textures and

[3]The sense of the positions and motions of our body parts.

weights and temperatures of objects are often determined by a combination of these two senses. Furthermore, the feeling of texture can be learned indirectly, as in the slipperiness of a patch of ice under the tires of a car.

Our sense of hearing provides many kinds of phenomenal perceptions. We can locate object spatially with it, determine the difference between the sound of a harpsichord and a piano playing the same notes. We know that the physical process involves compressions and rarefactions of air moving our eardrums, but we don't feel how this produces the rich variety of auditory experiences. Understanding how a record can contain that information via a single wavy line (the waveform) doesn't seem to help much. How does the sound of a guitar string vs. a piano correlate to this? Don't they just sound different, the guitar "guitarish" and the piano "pianoish"? It doesn't feel like we do a Fourier transform of a waveform at all. But we do: if you isolate the sounds of some of the harmonics that occur along with the fundamental tone of a plucked guitar string, and then listen to the guitar string again with these other notes in mind, you can now pick them out. These subtle supplementary notes are ususally phenomenally distinguished by saying "that sounds like a guitar" or "that sounds like a piano," but with some training the component parts can be picked out. Speech perception includes some more auditory phenomena: words seem to be clearly separated by gaps, gaps which do not in fact correspond to periods of silence. Instead, these gaps are grammatically defined. We also can distinguish between different tones of voice, between questions and statements, between sarcasm and fear.

Finally we turn to the phenomena of vision. Most of the visual properties of things we see seem to be in the objects themselves: colors, shapes, motions. Some can be recognized as arising from the interaction of light, the objects themselves, and our eyes, as in the glints of sunlight off of waves in a lake, but they still seem to be outside of us. All of the things we see are in the visual field, which seems to

be broad and detailed. However, when we test its resolving powers at the periphery, we find that we can distinguish only motion and some sense of light and dark.[4] In fact, we can only resolve things well two or three degrees from the center. This is concealed from us, is not part of our phenomenal view of the world, by the constant motion of our eyes. Perhaps the most important phenomenal impression from vision is that the world seems to be represented like a picture—like there is a picture in our head. This is a pervasive and misleading view, because though it seems to explain our phenomenal impressions it just couldn't be true. For who then is looking at the picture? And if there is something in there looking, what is in its head? This becomes an infinite regress of viewers within viewers, an unacceptable explanation. The place to break this regress is right at the beginning, with the first perception of a visual object. So although the phenomenology of vision seems like pictures in the head, there must be some better explanation.

What are the internal phenomena? These are thoughts, internal images, recollections. What are they made of? It seems impossible that they are formed of just the information received through the senses. How could mere sensory information, nerve impulses from the transducing organs of our senses, actually be the phenomenological items of thought? Thoughts often seem to be spoken of in terms of images: when we comprehend something we say "I see," and sometimes we mean it literally if the comprehension required visualization. However, even when people say they are visualizing something, even if it is the same thing, their visualizations differ widely. Dennett's example for this is comprehension of this sentence: "Yesterday my uncle fired his lawyer." ([Dennett, 1991], p. 56) Any two people hearing this sentence might have completely different images in mind but their understanding of

[4]Dennett provides an excellent example on page 53-54: take a card from a card deck without looking at it. Now focus on something directly in front of you, and slowly move the card from the edge of your vision toward the center. How close does it have to be before you can guess whether it is a face card? Its color? Its suit?

the sentence will probably not differ much at all, so images do not seem to essentially comprise understanding. Furthermore, how does one visualize "yesterday?" Dennett concludes that "comprehension . . . cannot be accounted for by the citation of accompanying phenomenology, but that does not mean that the phenomenology is not really there." (p. 57) Indeed, any model of thought that does not include the phenomenology will be incomplete. Internal thoughts and imaginings need not be composed of images, of course: we may also imagine the song "Happy Birthday" or the taste of oysters. That these recollections and imaginings are distinct from real experiences is clear, but what are they?

At last we turn to the affective members of our "phenomenal garden." Dennett gives us two prime examples of affective phenomena: sympathy and amusement. Imagine seeing your best friend forget their lines in the middle of a play. You might feel embarrassed for them, feel some of the same emotions they are feeling, become uncomfortable. Why do we have this kind of reaction? What is this phenomena? Imagine being an alien, and observing a group of animals making sounds at one another and then having some sort of convulsive reaction, one they seem helpless to stop, and furthermore seeking out circumstances which cause this phenomenon? Laughter seems purposeless, strange and unexplainable when viewed with detachment. Why do we have a sense of humor? What is it about certain things that makes them funny? We can explain pain fairly well—both its purpose in survival and its mechanism of action, but humor seems inexplicable.

These phenomena seem at the same time to be very familiar (what could we know better than our own mental phenomena?) and at the same time completely inexplicable. How could molecules or neurons be the smell of hot chocolate? This problem has been adressed by philosophers, who are concerned with the privileged access we have to our own experiences but not to others, and why this knowledge is different from he knowledge we have of things in the world outside; or who are

concerned with the intrinsic qualities of personal experiences, the qualia of pain or joy. As Dennett says, "finding a materialistic account that does justice to all these phenomena will not be easy." ([Dennett, 1991], p. 65) For some of these sensations we have enough understanding of the mechanisms which create them to make us challenge our introspective impressions, enough to give us hope that these phenomena can be explained.

Dennett's next move, after perusing the contents of the phenom, is to put forth a method for examining the contents more closely. When introspecting, he says, we make certain assumptions: that the contents of consciousness are clear to anyone willing to just look, that we share the same kinds of mental phenomena. Even though many people will go along with these assumptions, the conclusions drawn from the results always seem to cause controversy and disagreement. We must be fooling ourselves about one or the other of these assumptions, and Dennett thinks (rightly, I believe) that it is the reliability of introspection. Instead of actually observing our inner mental phenomena, perhaps we are really theorizing, inventing reasonable sounding explanations. As the experiment with peripheral vision shows, we often are surprised by the limitations of our senses.[5] Why are we surprised? Because we had some theory, perhaps never explicitly stated, about our peripheral vision. This theory comes not from actual observation but what seems reasonable. We don't feel like we see only a small amount of the world at a time, there don't seem to be gaps in our vision. Surely we would notice that everything we aren't looking at is not colored. If the visual field is like a picture, then all parts of it must be some color. We feel authoritatively sure about our conscious experience, to the extent that we are overconfident in our theories about it.

How can we study phenomenology without these kinds of errors? We have many

[5]I was when I tried the experiment, even though I 'knew' that the region of visual acuity subtends only a few degrees.

ways of studying human *behaviors* from an objective standpoint, but how can we study the seemingly inaccessible mental lives of others? This method must be neutral to avoid prejudging whether we can explain these phenomena, but it must be able to "do justice to the most private and ineffable subjective experiences, while never abandoning the methodological scruples of science." ([Dennett, 1991], p. 72) Dennett has invented such a method, called heterophenomenology. This method involves recording what people say and interpreting their utterances as intentional acts. This means that they "are to be interpreted as things the subjects *wanted to say,* of *propositions* they meant to *assert,* for instance, for various *reasons*." ([Dennett, 1991], p.76) All that they say is recorded with an uncritical eye, but also without assuming that they are telling the truth, or are even capable of intentional acts (they may be unconscious "zombies"). Dennett illustrates the use of this technique with the example of anthropologists studying a foreign tribe.

This tribe believes in a forest-god called Feenoman. If the anthropologists do not wish to convert to Feenomanism but still learn all they can about the tribe's religion, they may choose to be agnostic about his existence. They can ask the members of the tribe—Feenomonists—questions about Feenoman, gradually getting a picture of what he looks like and who he is. Some members of the tribe will disagree about certain aspects, like hair color or gender, and the anthropologists will have to settle these disputes, until finally a definitive picture emerges, a logical construct. It may contain contradictions, but the anthropologists don't care: Feenoman is, to them, an "intentional object" rather than a real thing, as opposed to the tribe's view. They remain objective about Feenoman's status. Now suppose that the anthropologists discovered a real person who was Feenoman, with blond hair, who ran around in the forest doing good deeds, but could not fly or teleport himself. This would probably disturb many of the Feenomonists, for it would question their religious beliefs about Feenoman. Indeed, unless the Feenoman discovered by the anthropologists was very

much like their description they would not believe that the discovered Feenoman was the cause of their beliefs.

Now apply this analogy to our mental phenomena. Using the method of heterophenomenology we can construct the definitive picture of our mental phenomena. Then if we could

> find real goings-on in people's brains that had *enough* of the 'defining' properties of the items that populate their heterophenomenological worlds, we could reasonably propose that we had discovered what they were *really* talking about—even if they initially resisted the identifications. ([Dennett, 1991], p. 85)

Further, any inconsistencies between people's descriptions of the heterophenomenological items and the "goings-on" we found could be justifiably called mistakes, even if sincere ones. Thus, heterophenomenology plus empirical investigation can give us an explanation of mental phenomena, an explanation it would be hard to argue against.[6]

## 4.2 Multiple Drafts vs. The Cartesian Theater

In this section is Dennett's first attack on the Cartesian Theater and the intuitions that lead to it, and his first statement of the Multiple Drafts theory. In it we will see how many ingrained habits of thought lead us to the mistaken view that there is someone in the head observing, or some place in the head, where the information of the senses is presented. This view is Cartesian materialism, "the view that there is a crucial finish line or boundary somewhere in the brain, marking a place where

[6]Of course people could insist that the goings-on and the phenomena accompanied each other but were not one and the same thing, but this sounds too much like epiphenomenalism, a doctrine without much explanatory power at all (See chapter 2).

the order of arrival equals the order of 'presentation' in experience because *what happens there* is what you are conscious of." ([Dennett, 1991], p. 107)

When we think of a conscious observer, we attribute to them a point of view. This is generally a useful attribution, for it explains what they can observe from where they are. We have no trouble explaining why the sound of fireworks arrives after the light, because we have a point of view specified and can refer to the speed of transmission of these to determine when they arrive at a certain point. However, when we try to locate the observer's point of view as a point inside them we run into problems. Although it seems as like there is some definable point (at least for ordinary time intervals) where one event arrives after another, when we look at smaller periods of time we have trouble:

> If the 'point' of view of the observer must be smeared over a rather large volume in the observer's brain, the observer's own subjective sense of sequence and simultaneity *must* be determined by something other than 'order of arrival,' since order of arrival is incompletely defined until the relevant destination is specified.[7] If A beats B to one finish line but B beats A to another, which result fixes subjective sequence in consciousness? ([Dennett, 1991], p. 108)

It seems like there must be a point where the sensory processes end up, after going through, say, the optic nerve and the primary visual cortex, and a point where the motor nerves that control the eye muscles trace inward toward, a definable place—a Cartesian Theater. This is the intuition we must fight against when trying to understand the mind. What, then, is a better metaphor and explanation for consciousness?

[7] I've found it useful to keep the neuropsychological evidence in mind when imagining these problems: different brain areas process different modalities, like sight and sound in the occipital & parietal and temporal lobes, respectively.

Dennett's replacement for the Cartesian Theater is the Multiple Drafts model:

> According to the Multiple Drafts model, all varieties of perception—indeed, all varieties of thought or mental activity—are accomplished in the brain by parallel, multitrack processes of interpretation and elaboration of sensory inputs. Information entering the nervous system is under continuous 'editorial revision.' ([Dennett, 1991], p. 111)

All sensory modalities do a large amount of "editorial revision" while performing their functions: for instance, the eyes (and head) are constantly moving, but this motion is edited out of phenomenal perception; when people see someone say "from left to right" (as in a movie) and hear "from reft to light" they will report hearing "from left to right." These discriminations occur in early sensory processing, and they do not have to be made again by some master interpreter watching the information come in. These interpretations, "content-fixations" in Dennett's terms, are locatable in specific brain areas and at specific times. This does not mean, however, that they determine what events become conscious. "It is always an open question whether any particular content thus discriminated will eventually appear as an element in conscious experience, and it is a confusion, as we shall see, to ask *when it becomes conscious*." ([Dennett, 1991], p. 113)

This stream of distributed content-fixations is rather like a narrative stream of consciousness, but is not quite the same: at any time there are multiple pieces of a narrative active and being edited. The narrative produced by the subject depends on when the stream is probed: if the stream is probed too late, like the next day, it will often be a reconstructed narrative. Probing too early may provide clues to how the contents are fixed, but in doing so will disturb the normal course of content-fixation. Thus there exist different narratives at different times, and their contents will depend on when the narrative is probed. This multiplicity of narrative fragments is like the

modern process of editing a text, hence the name "Multiple Drafts." When a book is in the process of being written, at any one time many versions of it could be circulating, especially since the advent of electronic mail: one version with the main editor, another earlier one with a friend of the author in Australia, the latest one on the author's computer. Which is the "real" text? They are all the real text, at different stages in the editing process. Often, the most important version of the text may be an earlier draft, while the final published draft is mainly archival: to those people for whom the text matters, the version they read was the "real" text, even though it wasn't the published version. Similarly, for the multiple stream of content-fixations, what the important contents are depends on which processes are using the information. Understanding this concept is crucial to understanding Dennett's model, so he provides some examples of how the Multiple Drafts model explains what have been some troubling experimental results in psychology experiments.

## 4.3 Orwellian vs. Stalinesque Revisions

When two spots are illuminated on a screen in rapid succession separated by as much as 4 degrees of visual angle, it appears as though there is one spot moving from the first location to the other. This effect is called phi. One of the most interesting forms of phi occurs when the two spots are different colors, like red and green. What seems to happen in this case is a red spot moves over and turns green about halfway between the two spots. What is troubling about this? The problem is, the intervening motion and color change don't really happen—they are invented by the brain. Furthermore, the intervening motion can't be inferred until after the green spot is seen. But if the green spot is perceived before the motion can be inferred, why does it seem like the order is (1) red spot (2) red-spot-turning-to-green-spot (3) green spot? How can we perceive the illusory motion, which depends on seeing

the green spot, as occurring before we see the green spot? Surely it follows that consciousness of the entire sequence must be delayed until after the green spot is perceived. No, says Dennett, this conclusion comes from the pervasive influence of the Cartesian Theater.

A thought experiment reveals how the Cartesian Theater influences how we explain this "editing" of experience. Imagine sitting in a park and seeing a long-haired woman jog by. Right after you see her, a memory of some other woman, with short hair and glasses, contaminates the memory so that you report having seen a long-haired woman with glasses. This is an Orwellian revision, after George Orwell who described a Ministry of Truth which rewrote history so that the actual history became inaccessible. On the other hand, your seeing the woman could have been contaminated on the upward path so that what you seemed to experience from the start was a long-haired woman with glasses. This would be a Stalinesque revision, because the experience you had was a sort of show trial, with false evidence and a predetermined conclusion. If there is a Cartesian Theater, one of these two accounts must be true.

By the Orwellian account, the actual woman was briefly experienced, then overwritten. By the Stalinesque account, the experience was falsified before it became conscious. The Multiple Drafts account, however, says that it is meaningless to ask questions about when particular contents got discriminated. These questions work fine for events in the world outside and at longer time intervals, but they do not apply all the way in: for if they they did, then the place where they apply is a Cartesian Theater, where all events are played on a screen for conscious viewing. The same problems with these two editorial processes come out in the color phi experiment. Both of them run into difficulties. The Stalinesque model has to posit a delay after seeing both dots, a delay in which the editor creates and inserts the intervening motion. This cannot be what happens, however, for if subjects are asked

to respond as soon as they see the red dot they respond at the minimum response time necessary to see the stimulus and move their finger. There is no time for a delay in response time. The Orwellian model posits a wiping out of the "real" experience (red followed by green) and a replacement with the illusory experience. The Stalinist can reply that the button push was an unconscious response, which the subject falsely insists they were conscious of. Both theories agree on all the facts of the matter, even about the place where the illusory content enters the narrative. All they disagree on is whether that place is pre-experiential or post-experiential. How can we distinguish between these two models?

Dennett says that we cannot, and neither can the theorists, and neither can the subject in the experiment.

> The two theories tell exactly the same story except for where they place a mythical Great Divide, a point in time (and hence a place in space) whose *fine-grained* location is nothing that subjects can help them locate, and whose location is also neutral with regard to all other features of their theories. This is a difference which makes no difference. ([Dennett, 1991], p.125)

Not only is it wrong to assume that there is a moment of consciousness, Dennett argues, but it is also wrong to assume that some sort of "film" of the illusory motion must be created at all. For whose benefit would this "film" exist? For the audience in the Cartesian Theater. It seems, phenomenologically, as if perceptual experiences are projected somewhere—sometimes outside of us, as in the case of stereo imaging;[8] sometimes inside us, as in color phi—but this does not mean that there is an actual place where these illusions are performed.

[8]In stereo imaging we often say things like "the sound of the piano is coming from right in front of me" —but of course it only seems that way, and no actual thing is projected out there.

> The representation of space in the brain does not always use space-in-the-brain to represent space, and the representation of time in the brain does not always use time-in-the-brain. "In short we distinguish representing from represented, vehicle from content." ([Dennett, 1991], p. 131)

This is not surprising: the brain must use some form of representation, one separate from the real world, pre-edited and pre-categorized. This is one frame of reference, the subjective frame, and "the space and time of what the representing represents is another," ([Dennett, 1991], p. 137) the external or objective frame. There is no projector in the brain, nor is there any "seems-projector." Just because our phenomenal experience feels like a continuous narrative shown in a theater, we do not believe this is really the case. There are no red spots in there, nor are there sounds. There are only contents which represent those objects, contents which are unconsciously discriminated and edited.

The Multiple Drafts model seems to deny the distinction between reality and appearance, a distinction that is preserved by the Cartesian Theater. Writing experiences down in memory becomes the criterion for consciousness, rather than projection to the Theater. Dennett describes this as "first-person operationalism"[9] because it "brusquely denies the possibility in principle of consciousness of a stimulus in the absence of the subject's belief in that consciousness." ([Dennett, 1991], p. 132) Another objection exists: Surely, one might think, there are two things in these cases. There is the judgement that motion seemed to happen (in the color phi case) and there is the thing that judgement was about—the "seeming-to-move" of the spot. Not so, says Dennett. There is only the judgement, the judgement is the seeming-to-move. The content must be discriminated only once, and since multiple content-discriminations are happening continuously, the narrative struc-

[9] Operationalism is the belief that if no difference can be discovered between two things, then there is no difference between them.

ture caused by a probe of these contents may vary from the order of events in the world. "There are no fixed facts about the stream of consciousness independent of particular probes," ([Dennett, 1991], p. 138) whether these probes come from within[10] or from without.

## 4.4 Objective and Subjective Time

The important idea that the representing of content in the brain is not the same as what is represented bears directly on our experience of time and timing of events in the world. Of course the seeming movement of color phi is not a case of backwards temporal projection, with the later frames being sent back to before the green spot is perceived. Nor is conscious perception of the whole event delayed until after the whole sequence is perceived. Instead, the brain makes several different discriminations in several different areas, and these are combined when probed into a narrative. Within the short span of the experiment, there is a "temporal control window" in which the temporal representations of events may be moved around, as long as the final content is accurate and comes in time to control the relevant behavior. ([Dennett, 1991], p. 151)

A narrative may be constructed out of order, as long as the narrative itself preserves the order in its content. For example, a movie may be filmed out of sequence, with the last scene filmed first. Furthermore, even the final structure of the narrative may be out of order but still preserve the relevant content, as in the phrase "B after A." ([Dennett, 1991], p. 149)

> What matters for the brain is not necessarily when individual representings happen in various parts of the brain (as long as they happen in time

[10]Dennett provides a good example of this kind of probe: the ability to count, "retrospectively in experience memory, the chimes of a clock which you only noticed was striking after four or five chimes." (p. 138)

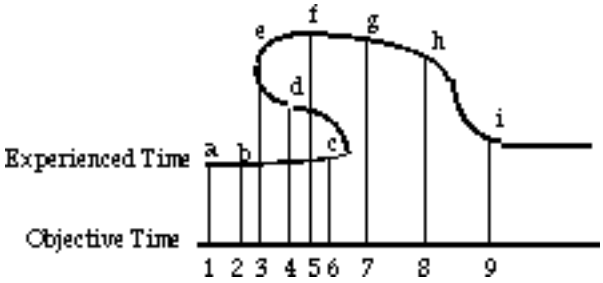


Figure 4.1: **Objective Time & Experienced Time** ([Dennett, 1991], p. 136)

> to control the things that need controlling!) but their *temporal content.* ([Dennett, 1991], p. 149)

As long as the information that tells the brain the true order of events in the world is accurate, it does not matter in which order it is discriminated. Certainly, of course, the timing of representings matters somewhat, because the events in the world that cause those representings do happen in a particular order, and the perceptions of them also happen in a particular order. But once the discrimination like "from right to left" has been made, that content can be sent elsewhere in the brain "in a temporally sloppy way, anywhere in the brain this information might be put to use." ([Dennett, 1991], p. 150)

What matters to the brain as an interpreter of events in the world and as a controller of behavior in the world may vary from the objective facts about sequences of events in the world, as shown in the two sequences in 4.1. When asked about their perceptions of a series of events, subjects may tell a story that sounds like the upper sequence. Since the experimenters know that events "e" and "d" came before event "c" they may be puzzled by the subjects' reports, puzzled enough to suggest that perception of events is somehow projected back in time, puzzled enough to suggest that their results require reworking physics or giving up on a materialistic explanation of conscious experience. There is no need for such extreme measures,

however, if we give up the Cartesian view of the mind. These results are explainable if we instead adopt something like the Multiple Drafts model, which recognizes that there are many parallel processes occurring simultaneously in many different regions of the brain, and that a subject's narrative report of their experience may be in a different order than would seem possible, because of the interpretations the brain produces and because the content fixations in the brain take place at different places and at different relative speeds.

## 4.5 Evolution, Language and the Architecture of the Human Mind

How did consciousness arise?[11] It must have evolved from processes that weren't conscious, so an explanation that can accommodate that transition should help us understand some of its current features. Dennett paraphrases Valentino Braitenberg's ideas from the book *Vehicles* [Braitenberg, 1984] to point out why starting from the bottom can help explain higher levels. Braitenberg puts forth the law of uphill analysis and downhill synthesis: understanding a complex organism (or machine) from its behavior as if it were a "black box" is quite difficult. On the other hand, if the design of this organism (or machine) is known and all of its parts are comprehensible, then it is much easier to imagine its behavior.

The first living things on earth had to have the ability to make a few different discriminations in order to survive and reproduce themselves. Of course they were not conscious of these discriminations, and attributing reasons for their behavior in terms of intentions is probably ridiculous. Nevertheless, even at this level things which behave in their own interests can be described as having a sort of point of view.

[11] Many of the sections of Dennett's book I will summarize here are highly speculative, or are meant merely to make the idea of a conscious brain more plausible, to make imagining a mechanistic explanation easier, and so should be interpreted in that light.

To succeed they had to do certain things: firstly, they had to distinguish between inside and outside, between self and everything else. This can be accomplished purely mechanically, as in the case of the immune system, where the difference between "good" or "ours" and "bad" or "outsider" is determined by particular molecular configurations.

The next foundational point Dennett makes is about the operation of "Mother Nature" or natural selection. Design is not laid out intelligently, so some new developments may have unplanned side effects. These side effects may confer advantages, or they may not. Sometimes the side effects of a development may be more important for the organism than the original effects, and so the development plays more than one role. Multiple functions are normal in nature as opposed to human design, and make organisms more compact and flexible.

After a point of view is established, organisms must develop a way of anticipating the future if they don't adapt the strategy of a barnacle. For these organisms which move themselves away from the "bad" and toward the "good" the important problem is "Now what do I do?" ([Dennett, 1991], p. 177) Having some mechanism which makes the organism either approach or avoid a stimulus gives a creature a sense of the immediate future. One such primitive mechanism is touch, but even better would be some way of avoiding bad things and approaching good ones a little in advance, by picking up some kind of regularity in the world. An existing example is a "duck" mechanism, for suddenly looming objects, such that the stimulus "looming object" is wired to the "duck" response. Another might be a vertical symmetry detector [Braitenberg, 1984] for detecting other organisms looking toward you. This detector might often fire mistakenly, but its advantages would outweigh its deficiencies by making the organism cautious. When this detector fires, it would turn on other parts of the nervous system, alerting the "friend or foe" detectors and perhaps others—a period of heightened awareness. Once such an alerting mechanism developed, it

might turn out to be useful to have it on all the time, to keep the rest of the nervous system generally "aware" rather than generally "on automatic." "Regular vigilance gradually turned into regula *exploration,* and a new behavioral strategy began to evolve: the strategy of acquiring information 'for its own sake' just in case it might prove valuable someday." ([Dennett, 1991], p. 181) The information gathering devices, however, still retained some of the "friend or foe or food" overtones of their original designs.

The next development, and perhaps the most important, is the development of plasticity. Plasticity is the ability to change the way the nervous system is "wired," and allows much faster development than genetic evolution. With genetic evolution, "Good Tricks"[12] are very slow to develop because very few individuals may accidentally be born with them. Plasticity allows members of a species which are fairly close to having a Good Trick to replicate, while making only those members which are unable to get to that state fail. This speeds the acquisition of such tricks, since phenotypic postnatal development allows them to be slightly "off the mark" and still acquire the trick. Plasticity seems something like Lamarkian evolution, since postnatal learning plays a part, but it is really still genetic evolution. The difference is that it becomes easier for those close to having the Good Trick to survive, not just those born with the Good Trick.

Plasticity is one mechanism needed to explain how consciousness could arise, but it is not enough. We need to explain how humans could become truly excellent "future-producers," how they could think ahead, develop long-term plans, avoid "ruts," and deal with novel situations. Two problems arise here: what to think about next, and how to represent these things. The problem of what to think about next may depend on a process like pandemonium, where individual subsystems com-

[12] A Good Trick is any development which confers a survival advantage to its possessor. ([Dennett, 1991], p. 184)

pete for dominance. "We should not expect there to have been a convenient captain already at hand (what would he have been doing up till then?), so conflicts between volunteers [recruited by the alerting mechanism] had to sort themselves out without any higher executive." ([Dennett, 1991], p. 188) Dennett believes that this process is not enough to explain our abilities: just as the original alerting response was triggered by the environment but eventually became constantly triggered from within, there might be some pressure to solve the problem of what to think about next from within. The solution to this problem, and to the problem of representation, may require something more like a stream of consciousness, something which could control lower-level processes without being distracted by the environment. Dennett's proposal for explaining this thing relies on several key ideas: autostimulation and the development of language, memes and cultural evolution, and the idea of a virtual machine.

The ancestors of humans, early primates, probably had some way of communicating with the other members of their group. This is not such a wild supposition: some kinds of monkeys today have distinctly different vocalizations for signaling the approaches of predators from the sky (the "eagle" cry) or from the ground (the "baboon" cry). More advanced speech acts seem plausible, as in communicating "there is food here." Suppose as well that these ancestors developed ways of soliciting these speech acts—in effect, asking questions ("Is there food here?"). Now imagine that one day one of these primates asked a question when no other was around, and this auditory stimulus provoked the proper response—an answer to its own question. This step requires positing that the mechanism for answering questions was not properly "wired" to the mechanism for answering them. This autostimulation might be a Good Trick, a new way of soliciting useful information, such a Good Trick that it became quickly adopted and refined, perhaps into subvocal speech. It would be a much slower and clumsier process than existing unconscious cognitive

processes, it would be linear in form (just like the speaking and responding it was evolved from), and it would be limited to the kinds of questions and answers the animal already knew how to produce. This kind of autostimulation might also occur visually, through drawing pictures: seeing two parallel tracks idly drawn might bring to mind the banks of a river, reminding the hominid to bring his fish-gutting stick. This kind of visualization could become automatic, so that no pictures would be needed. Are these kinds of novel autostimulations plausible? Human split-brain patients seem to perform them in order to overcome the lost connections of their corpus callosum: where data from one hemisphere is insufficient to solve a problem, the other may be provoked to action by speaking out loud, or by stimulating pain sensors which are connected ipsilaterally rather than contralaterally.

Just as phenotypic plasticity depends on genetic variation, cultural evolution and the transmission of memes depends on phenotypic plasticity. The existence of this plasticity allows organisms to learn Good Tricks postnatally, and if these tricks can be preserved and trasmitted over time, a new kind of evolution can take place—cultural evolution. Cultural evolution and the contribution of memes both depend on brains that understand language, a process that humans do very swiftly and easily. This is probably because the ability to use language is such a Good Trick that those who were best at it passed on their genes, and those who could not learn language easily were at a tremendous disadvantage.

What are memes? Memes are Richard Dawkins' word for replicators in the form of ideas (from [Hofstadter and Dennett, 1981]. They are meant to be thought of in the same way as genes. They are the smallest units of cultural transmission, units which go from brain to brain just as genes go from body to body, units which replicate themselves, units with differential fitness values. Examples include "tunes, ideas, catch-phrases, clothes fashions, ways of making pots or of building arches."

(in [Hofstadter and Dennett, 1981], p. 143)[13] Memes may be useful, entirely useless, or even dangerous, but they depend only on replication to survive, not on the effect they have on the brain which they inhabit. Thus there is not necessarily a connection between the meme's fitness and its usefulness to the brains it lives in—but a meme which caused organisms to jump off cliffs would have the same kind of fitness as a gene which caused the same behavior. It would probably not survive long. Memes like the concept of a spoked wheel or a fishing hook could obviously confer survival advantages, and when the organism can communicate them through language they will replicate to all others who can understand language. The restructuring of the brain caused by language allows all of the cultural memes to enter it. This ability comes from the habits of autostimulation, which depends on neural plasticity, which depends on the phenotype, which in turn depends on the genotype. Thus, once an organism has evolved up to the point of having a language it can rapidly advance by taking advantage of cultural memes which do not need to be reinvented. Memes are important for Dennett because he believes that the "human mind is itself an artifact created when memes restructure a human brain in order to make it a better habitat for memes." ([Dennett, 1991], p. 207) This restructuring makes the brain vastly more powerful, able to use memes created over centuries and able to use other memes to combine them together. Describing the effect of this change requires a new level of description, one analogous to the description of software on a computer. The thing thus described is a virtual machine, and that is just what Dennett believes consciousness to be. Here is his own summary:

> Human consciousness is *itself* a huge complex of memes (or more exactly, meme-effects in brains) that can best be understood as the operation of a *"von Neumannesque"* virtual machine *implemented* in the *parallel*

[13] Page numbers from the excerpt of *The Selfish Gene* reprinted in [Hofstadter and Dennett, 1981].

> *architecture* of a brain that was not designed for any such activities. The powers of this *virtual machine* vastly enhance the underlying powers of the organic *hardware* on which it runs, but at the same time many of its most curious features. and especially its limitations, can be explained as the byproducts of the *kludges* that make possible this curious but effective reuse of an existing organ for novel purposes. ([Dennett, 1991], p. 210)

Much of the jargon in this paragraph has already been explained in chapter two: "von Neumannesque" is analogous to Turing-machinesque, i.e. a serial, one item at a time method of processing. The terms "virtual machine" and "kludges," however, require some explanation. A virtual machine in computer science is a temporary set of rules that acts on a particular kind of input and produces a particular kind of output. For example, Microsoft Word™ is a virtual machine that takes in input from a keyboard and mouse and puts out text ordered in particular ways. The same virtual machine can be implemented on different kinds of hardware: Microsoft Word for an IBM looks and acts just like Microsoft Word on a Macintosh or an Amiga or any computer capable of handling the inputs, outputs and rules required by the program. The same underlying hardware, on the other hand, can run many different kinds of virtual machines: a chess machine, a calculator, a CD player. Of course consciousness is not a program, for many reasons: the way computers load programs and the way humans become conscious are markedly different, human brains do not have a fixed machine language like computers, different brains can have vastly different systems of neural interconnection whereas computers must be exactly the same, etc. There are good reasons for this sort of analogy, however. Understanding consciousness at the "program" level rather than the neural level should be much easier. No-one could understand computer programs at the individual circuit level even given all of the microsettings of the computer's memory and CPU.

The term "kludge" means an after-the-fact modification of software, often inelegant, inserted to make the software actually work. The analogy with serial consciousness is between the processes that a serial device needs (like fast memory creation, access and retreival) and the simultaneuous parallel processors that the device is implemented on. The representativeness and availability heuristics dicussed in the last chapter are evidence of kludges: ways of getting at information that work pretty well, and quickly, but are not algorithmic. The processes humans have evolved for the interface between the virtual machine of serial conscious thought and the hardware that implements it are Good Tricks. Two of them are the rehearsal and association processes for getting information into the brain, and the converging association process for recovering information. These are quite different processes from those used by von Neumannesque computers, and probably far less elegant than the simple instructions for retrieving memories from specific addresses, but they work much more quickly and seem to be able to handle the knowledge problem far better than computers can.

How could this be any sort of explanation of consciousness? After all, computers are entirely unconscious, and they are intrinsically serial processors. Dennett's answer is this:

> The von Neumann machine, by being wired up from the outset that way, with maximally efficient informational links, didn't have to become the object of its own elaborate perceptual systems. The workings of the Joycean machine,[14] on the other hand, are just as "visible" and "audible" to it as any of the things in the external world that it is designed to perceive—for the simple reason that they have much of the same perceptual machinery focused on them. ([Dennett, 1991], p. 225-

[14]Dennett's term for serial consciousness. It comes from "the meandering sequence of conscious mental events famously depicted by James Joyce in his novels." ([Dennett, 1991], p. 214)

226)

This model depends somewhat on the machinery that implements it, thereby giving it more plausibility than many cognitive explanations at this level. It is an improvement in some ways over the other models I have presented so far in that it takes advantage of their most plausible ideas and uses the facts about the brain to help explain not only how the physical architecture of the brain works but also why our phenomenological experience of consciousness is as it is. Dennett has been brave enough to speculate from the consequences of the findings of neuroscience, neuropsychology, and cognitive psychology to an explanation of how consciousness could come about, work in the ways that it works, and most importantly feel the way it feels. Before proceeding on to Dennett's discussion of the problem of qualia and some discussion of problems with Dennett's book, I will present his Thumbnail Sketch of the Multiple Drafts model:

> There is no single, definitive "stream of consciousness," because there is no central Headquarters, no Cartesian Theater where "it all comes together" for the perusal of a Central Meaner. Instead of such a single stream (however wide), there are multiple channels in which specialist circuits try, in parallel pandemoniums, to do their various things, creating Multiple Drafts as they go. Most of these fragmentary drafts of "narrative" play short-lived roles in the modulation of current activity but some get promoted to further functional roles, in swift succession, by the activity of a virtual machine in the brain. The seriality of this machine (its "von Neumannesque" character) is not a "hard-wired" design feature, but rather the upshot of a succession of coalitions of these specialists.
>
> The basic specialists are part of our animal heritage. They were not de-

> veloped to perform peculiarly human actions, such as reading and writing, but ducking, predator-avoiding, face-recognizing, grasping, throwing, berry-picking, and other essential tasks. They are often opportunistically enlisted in new roles, for which their native talents more or less suit them. The result is not bedlam only because the trends that are imposed on all this activity are themselves the product of design. Some of this design is innate, and is shared with other animals. But it is augmented, and sometimes even overwhelmed in importance, by microhabits of thought that are developed in the individual, partly idiosyncratic results of self-exploration and partly the predesigned gifts of culture. Thousands of memes, mostly borne by language, but also by wordless "images" and other data structures, take up residence in an individual brain, shaping its tendencies and thereby turning it into a mind. ([Dennett, 1991], pp. 253-254)

Is this really a theory of consciousness? Could something have all of the features and properties described above, and yet still not be conscious—a zombie? The argument for this position usually says something like "I could imagine something with all of those functional parts that wasn't conscious; therefore you haven't fully explained consciousness or even gotten close to describing its special property." Dennett's reply is like his replies to Jackson's knowledge argument and Nagel's "What is it like to be a bat"[15] arguments: Could you really imagine such a thing? This is a common failing in philosophy, to mistake a failure of imagination or an ill-conceived stretch of imagination for a failure in the theory in question. Such proofs from imaginability, in spite of the weight of evidence against them, have a curious tenacity—in cognitive psychology it is known as the confirmation bias.[16] In the next section I will present

[15] See chapter 2 on property dualism.

[16] The confirmation bias refers to people's unwillingness to give up a theory: even when the

some of Dennett's ideas on qualia, the supposedly unique and incorrigable items of subjective experience.

## 4.6 Qualia Disqualified

Our experience of colors is a favorite philosophical example of a quale:[17] science has shown us that colors are not out there in the world because all that is is out in the world are electromagnetic radiations, some of which we are sensitive to. Therefore, colors must be in the eye and brain of the beholder—purely subjectively. When I see red, there is the 650-750 nanometer wavelength radiation "out there," and then there is the "redness" that I see "in here": the way red looks to me. Before going on to the arguments meant to prove the existence of the mysterious properties that such experiences have,[18] Dennett puts forth some alternative explanations for the phenomena we experience.

Colors, according to Dennett, "*are* properties 'out there' after all." ([Dennett, 1991], p. 373) The idea of red is just a discriminative state that has the content: red. When we compare two colors in our "mind's eye" we are just retreiving the contents of memories, memories which have the two color contents. This is analogous to how a computer (which could discriminate colors) would compare two colors: it would check its memory, and determine from the contents (which might be labeled "red 223" and "blue 410" or some such) whether the colors were the same. Just what are these colors, then? Are they just the spectral surface reflectance of objects? No, for the reason that humans call some things "red" which are quite different as

evidence that caused them to have the theory is demonstrably false, and the theory leads to implications which are false, nevertheless people will cling to it.

[17]Singular of qualia.

[18]The "inverted qualia" argument and "Mary the color scientist": the first falls to Dennett's exposition of reactive dispositions and neurophysiology, and the second falls because it implies that qualia are epiphenomenal.

far as scientific measurements of the wavelengths that reflect off them tell us. Human color discrimination is more complex than that, as experiments with lighting conditions and context tell us. Nonetheless, our subjective colors are just content discriminations.

Dennett argues that color vision evolved along with things in the environment, in order to discriminate things in the environment. Flowers, for instance, evolved colors that make them bulls-eyes for creatures with ultraviolet-sensing visual systems, just as creatures with such visual systems (e.g. bees) evolved to sense flowers. Why to apples turn red when they ripen? There is a chemical explanation for the color change, but this fails to capture the reason such a color change exists: to make them highly salient to apple- eaters who will spread their seeds. The fact that they turn red (as opposed to some other color) has much to do with the kinds of visual systems that saw them—that a photopigment that responds to red was available among fructivores. Our visual system evolved not to detect differences in incident wavelengths of light but to pick out important features in our environment. What we call "red" may vary widely, as long as it picks out the important things.

Furthermore, there are reasons we prefer some colors to others, fear some sounds and enjoy others, reasons which are explainable not by intrinsic subjective qualities but by evolutionary pressures and intra-species variation. The fact that we have particular value-states in us when we experience things is not unexplainable by science, and is not due to our own particular qualia, but is due to the original purposes of our sensory systems. For our ancestors, sensory systems were just things that made us run away from some things and approach others, and these original "wirings" still may have effects on our sensory experiences, giving them an affectual component as well as an informative component. Our experiences are "colored" by these original connections, even though they have little use for us today in our artificially created and colored environment. Not just the original connections alter

the affective components of experiences, of course. Each person's private series of interactions with their environment can cause them to have different reactions to various colors, tastes etc. because of events in their past. For example, for me pea-green might be a pleasant color, reminding me of my grandfather's excellent split-pea soup, but for someone else it might bring up memories of a playground bully's shirt. These associations need not be accessible to have effects—the connections may be entirely unconscious.

What about the idea that the red you see and the red I see are different? To answer this question, we must ask another: How could we tell? If we identified all the same things as red together, both had the same sorts of machinery in our eyes and brains, and all other physical aspects of color vision, what could possibly be different? If the qualophile insists that there still could be a difference, they must point out that difference. And if it turns out that they can only respond that the difference is "that special quality I personally feel when I see red, independent of everything else" then they are dodging the question. The only escape from the sort of explanations presented so far is to say that qualia are epiphenomenal—a philosophical position examined in the chapter 2 and shown to be lacking.

## 4.7 Imagining Consciousness

Has Dennett succeeded in explaining consciousness? Some might say his explanation leaves something out, it explains away consciousness. Dennett replies that successful explanations always leave something out—otherwise they wouldn't be explanations. For example, an answer to the question "What is the difference between gold and silver?" might be to say that they have different numbers of subatomic particles. Has this explained away the goldness of gold?

> Only a theory that explained conscious events in terms of unconscious

> events could explain consciousness at all. If your model of how pain is a product of brain activity still has a box in it labeled "pain," you haven't yet begun to explain what pain is, and if your model of consciousness carries along nicely until the magic moment when you have to say "then a miracle occurs" you haven't begun to explain what consciousness is. ([Dennett, 1991], pp. 454-455)

If explanation of consciousness in terms of unconscious things is wrong, if it leaves something out, then surely the explanation of living things in terms of things that aren't alive is also no good. Has Dennett succeeded in explaining consciousness? In his final summary, Dennett says:

> My explanation of consciousness is far from complete. . . All I have done, really, is to replace one family of metaphors and images with another, trading in the Theater, the Witness, the Central Meaner, the Figment, for Software, virtual machines, Multiple Drafts, a Pandemonium of Homunculi. It's just a war of metaphors, you say—but metaphors are not "just" metaphors; metaphors are the tools of thought. No one can think about consciousness without them, so it is important to equip yourself with the best set of tools available. Look what we have built with our tools. Could you have imagined it without them? ([Dennett, 1991], p. 455)

In my summary of Dennett I have had to leave out many interesting discussions, including the notion of the self as a center of narrative gravity, much of his discussions of internal representation, much of his discussion of intentional speech, and some interesting ideas from artificial intelligence. Even so, I believe the ideas I have presented are powerful and go a long way toward explaining how we could get mind out of meat. In the next section I will present some objections to Dennett's views,

and then proceed to my conclusions.

## 4.8 Problems with Dennett

Thomas Nagel, in his review of *Consciousness Explained* [Nagel, 1991], accuses Dennett of leaving out mental events: "A theory of consciousness that doesn't include mental events among the data is like a book about Picasso that doesn't mention his paintings." This problem seems to have been addressed by Dennett in the last section of the book, but Nagel wants a subjective theory of consciousness rather than an objective one. Nagel disagrees with Dennett's method of interpreting mental events, a method which Dennett agrees depends on people's own descriptions of their phenomenology. Heterophenomenology, Dennett's method, is no good according to Nagel because it would allow "a sufficiently complex and animated but subjectively unconscious zombie" to be conscious. Finally, Nagel disagrees with Dennett's ideas about what human consciousness is: by Dennett's description, a human baby can't be conscious because it has no virtual machine yet, no language, no self-perception in the way that adults have a self-perception.

Nagel's problems with Dennett all seem to stem from his belief that there is something intrinsically inexplicable about consciousness from an objective, third-person, scientific viewpoint. Briefly, Nagel is a property dualist, and adamant about it. That he believes that zombies are possible (in the philosophical sense) points this out: there could be something that is in no way distinguishable from a conscious human and yet is nonetheless unconscious, lacking "subjective consciousness." Nagel's point about babies seems troubling, but what is wrong with supposing that babies (not to mention dogs) have a limited sort of consciousness, a biological sense of self rather than the sort of narratively defined, meme-constituted self that adults have? I wonder if Nagel would believe that a human fetus is conscious, or at what stage

he believes it becomes conscious.

Ned Block's [Block, 1993] problems with Dennett involve Dennett's supposed conflations of several distinct kinds of consciousness: phenomenal consciousness, self-consciousness, and access consciousness. Phenomenal consciousness is distinguished from the other, cognitive kinds of consciousness by its content—it contains experiential content, which may overlap with the other kinds of content but is different. The contents of self-consciousness are thoughts about oneself; the contents of reflective consciousness are thoughts about one's own mental states; and the contents of access consciousness are semantic or representational thoughts. "A state is access conscious if its content is inferentially promiscuous (in Stephen Stich's sense), i.e., freely available as a premise in reasoning; *and* if its content is available for *rational* control of action and speech." ([Block, 1993], p. 182) Because Dennett conflates these three distinct forms of consciousness, Block argues, his argument about the Orwellian/Stalinesque distinction fails. Although Block says that "no currently available neurophysiological or computational concepts are capable of explaining what it is like to be in a phenomenally conscious state, e.g., to have a pain or to see something red" ([Block, 1993], p. 182) he believes that someday there might be—perhaps Crick and Koch's 40 Hz oscillations. If we had such a method, we could distinguish between the Orwellian and Stalinesque theories, because in the Orwellian case there would be a brief flicker of phenomenal consciousness, detectable by an outside observer. Block has one more problem with Dennett, a problem which seems to reveal a commitment to a Cartesian viewer:

> "The main part of the theory is the widely held view that the mind is composed of numerous semi-autonomous agencies competing for control. Much of the content of the theory concerns the original functions of these agencies and their mode of interaction and organization. *I see little in the theory that is about consciousness in any sense narrower than the*

*whole mind."* ([Block, 1993], p. 186 [emphasis mine])

Block's phenomenal consciousness shares the same problems as Nagel's subjective consciousness, except that he believes that it is in principle a discoverable thing. He believes it to be something special, over and above the content-fixations of our parallel systems—the phenomenal part of those contents. He refuses to accept the reduction given by Dennett in the chapter on qualia. Block's objection to the Orwellian/Stalinesque argument fails because he does not fully understand the Multiple Drafts argument. Both George Mandler and myself had the same confusion, which stems from the tenacious hold of the Cartesian Theater: Surely there must be a fact of the matter! Surely the contents of consciousness can be said to exist or not exist independent of our probes (or the subject's own probes)! I will discuss this problem later, in the introduction to my experiment. Block's problem with the breadth of Dennett's theory, that it does not zero in on which part is conscious, is strange: surely he would admit that there is no conscious observer in the brain, so why is it wrong to explain consciousness in terms of all of the parts whose functioning creates it? In short, I believe Block is still using the metaphors of the Cartesian Theater to try to understand mental events, metaphors which often lead to misleading intuitions and hence false conclusions from those intuitions.

George Mandler has problems with Dennett because he argues "from the inside out" and in doing so neglects much of the work that has been done "from the outside in" in cognitive psychology. ([Mandler, 1993], p. 335) For example, Dennett seems to believe that all learning is conscious, "a position that leaves out large chunks of motor learning, and assumes that learning to play bridge, to ride a bicycle, and to drive a car are cut of the same cloth." ([Mandler, 1993], p. 337) Dennett ignores the important distinctions between implicit and explicit learning and memory processes. Mandler also accuses Dennett of understating the effects of unconscious processing, of not fully discussing why a serial consciousness is useful,

and of not spending enough effort discussing the functions of consciousness. Furthermore Mandler thinks Dennett's statements about evolution are really about cultural evolution, and Dennett "makes too much of the memes," the units of cultural evolution. ([Mandler, 1993], p. 337) If consciousness is not hardwired but instead entirely a social product (composed of memes), then why is there such universality of human consciousness across societies? This universality should lead us to suspect genetic factors rather than social ones. Mandler also discusses the Orwell/Stalin distinction and finds Dennett's conclusion lacking. Finally, Mandler states his own positive account of consciousness (one much like Norman and Shallices' discussed in 2.2.3) as a more passive function, "a useful gatekeeper that on the one hand permits limited entry and on the other hand, by feedback of of activation and differential selection, affects subsequent conscious contents." ([Mandler, 1993], p. 338)

I believe many of Mandler's criticisms are right on: Dennett misses much of cognitive psychology, especially in the areas of unconscious processing and implicit vs. explicit learning and memory. Dennett goes partway toward explaining why consciousness should be serial, but most of his arguments are to assert the seriality of consciousness rather than explain why it is useful. Dennett's over-dependence on culture and memes for the shape of human consciousness bothered me as well: he seems to be ignorant of many of the neuropsychological data on language and semantic processing which indicate more hardwiring than the meme-cultural evolution thesis implies. On the other hand, Dennett makes a valiant effort at challenging the notions we derive from our phenomenology, an effort few psychologists have made and at which none have succeeded in anything like the way Dennett has.

My own problems with Dennett's theory follow Mandler's quite closely. Dennett made a good start from phenomenology to heterophenomenology to the Multiple Drafts theory. That much of the brain operates more like a coalition of homunculi than like a Cartesian viewer seems quite clear, and Dennett does a good job of

replacing many of the common, misleading metaphors with creative new metaphors that seem to lead in the right directions. Dennett's insights into the evolutionary reasons for many of our brain's features are good—too few people have explained any part of consciousness teleologically,[19] much less with plausible teleological insights. I believe, however, that Dennett stops the evolution too soon, and therefore depends too much on his virtual machine and mind as program metaphors. Some processes must depend on memes reorganization of how we think, but much more of the machinery than Dennett appears to believe is built in, hardwired in some way.

## 4.9 How the Brain Works: Explaining Consciousness

Unless we want to be unreasonable and insist on an epiphenomenal dual world for mental events, we must believe that somehow the workings of the brain produce consciousness. We must believe that our memories of high school, our ability to write a well-reasoned paper, our experience of eating strawberry ice cream, our ability to do arithmatic, to speak, to learn, to feel pain, to laugh and cry—all of these and more—depend in some way on a bunch of neurons releasing neurotransmitters at each other. This is hard to imagine, so hard that many people are driven to believe that it is impossible in some way, either metaphysically or epistomologically, and so they conclude that consciousness will remain forever a mystery. I reject those notions because I am still curious and optimistic, because I believe we can understand how we get mind out of meat through the application of our tools of thought. These tools include scientific experimentation of many kinds as well as conceptual tools: metaphors and analogies. By applying the right metaphors to scientific data we acheive understanding of the world; by building models we come

[19]Relating to purpose or design, especially natural purpose or design. Thus the explanations of biological features mention their goals.

to see the implications of our theories and thus gain deeper insights.

To build a theory about as large and complex a system as a brain we need a great deal of knowledge: knowledge about what it is made of, how it is put together, what each of its parts do, how those parts function together to produce the behavior we observe, why it produces the behavior we observe. An analogy may help here, on a much simpler system, like an airplane.[20] Imagine we had no insight into the design of an airplane, but were curious about what it was, and how a machine could fly. What could we learn that could answer these questions? We can learn about the materials that compose an airplane and how they are attached to one another: the properties of metal, rubber, plastic, glass, etc. and how screws, bolts, welds, clamps and such work. We can learn about the larger groups of those materials, and their functions: the elevators, turbines, ailerons, camshafts, exhaust manifold, wheels, etc. We could learn about the capacities of the airplane: how fast it can go, how much fuel it uses, how long it takes to get off the ground, etc. We might also need to learn about the reasons such a thing exists, what it is for in its environment, its teleology. Given all this, we would know what an airplane was, and have no problem explaining how it could fly. We would have successfully explained the mechanism and purpose and behavior of airplanes.

Explaining life once seemed as impossible as explaining consciousness seems today. How brute, dead matter could be put together in such a way that it becomes alive seemed impossible in principle: without the vital force, *élan vital*, nothing could be alive. The mechanisms of life are complex and require a great deal of visualizing and knowledge to understand, yet they are still mechanisms. Can we pull off such a feat with consciousness, with the mechanisms of the brain? Yes, by accomplishing analogous feats of visualization and acquiring analogous quanti-

[20] I am not comparing the brain to an airplane here, but instead trying to point out levels of analysis and the kinds of things each one can teach. I am fully aware of how much more complex a system the brain is, but I think these are valid points—bear with me.

ties of information. Biochemistry is far from complete. Many mechanisms are not fully understood, perhaps a larger portion than are understood, but we still do not believe that understanding is impossible. It is merely difficult, and new ways of approaching problems are helping us understand enzyme action and protein structure and the myriad other complexities of living systems, new techniques like x-ray crytallography and computer imaging.

What has this survey of knowledge and theories shown about consciousness? From neuroscience we have learned about the physical architecture of the brain. Glial cells, axon bundles, cortex, neurons of many kinds, neurotransmitters: these are the component parts of the brain, and facts about them—about how they are put together, how they interact, where the connections between them lead—are crucial to an understanding of the functioning of the brain as a whole. The speed of neurotransmission is an important theoretical constraint, one of the main reasons we believe the brain must be a parallel processor. The types of neuronal circuits give us some clues about how timing, motor control, and other such processes could be explained. Different neurotransmitter systems can help us individuate mechanisms for emotions, memories, arousal, depression, and learning. Neuroscience also studies larger systems, such as those that process visual input. That different brain areas process different aspects of visual inputs is an important fact for explaining conscious visual experiences, and if generalizable to other sensory functions this sort of parallel mechanism is important for explaining all conscious sensory experiences. Certainly, then, facts about the components of the brain should constrain our explanation of consciousness, but they are not all there is to know. Whatever structures we believe produce consciousness must ultimately reduce to neural-level details, but arguing from the neural level about consciousness is like arguing from the physical level about neurons: the difference of scale is too great, and we are none of us Laplacian gods who can know all the physical information to such a degree that we can determine

the behavior of a neuron from it. We need to theorize at a larger scale.

Cognitive neuropsychology has taught us much about the functional architecture of the brain, what the groups of neurons described by neuroscience do. From neuropsychology we have learned that the many functions of the brain are accomplished by many different distinct brain areas, distributed throughout the cortex. Processing of visual information, for example, occurs in specific parts of the occipital lobe of the brain, while auditory processing occurs in the temporal lobes. Furthermore, memories specific to each modality are stored in the same areas of cortex that do the modal processing: content is stored where it is discriminated. We also know that language processing takes place in the left hemisphere of the brain, with separate parts responsible for separate linguistic abilities. Neuropsychological findings have inspired productive theorizing about consciousness, especially the concept of modules developed by Fodor and refined by Norman and Shallice and Moscovitch and Umilta. Their hierarchical model of modular functions uses Fodor's excellent characterization of low level processing and adds some more conceivable ideas about higher level functions. A more mechanistic picture emerges, helping explain the powers of Fodor's central systems by combinations of lower and higher level modules. Groups of unconscious homunculi—each of which can be said to "care about" different areas of thought—are activated at different times, contributing their contents to the stream of consciousness. How these homunculi contribute to the serial stream of consciousness is unknown, but processing that takes place in the frontal lobes of the brain appears to be crucial.

Cognitive psychology shows us some of the limitations and idiosyncracies of human thought. From these it infers what kinds of information processing go on to create mind. Since it works at a functional level it is able to describe more abstract features of the brain's architecture, features that could not be deduced from low level physical knowledge. Cognitive psychology explains many of the conscious, serial

processes that form our conscious experience of the world, like memory, perception, attention, learning and reasoning. We have learned that our perception of the world comes to us precategorized by unconscious processes, that our speech production system is implicated in learning and working memory, that humans are not logical, rational thinkers but instead use much faster and less reliable heuristics. From information science we have learned some of the limitations of paradigmatic serial processors like computers, limitations which lead us to believe that much cognitive processing cannot be produced serially. Connectionist or PDP models seem to mimic some of the ways human brains work far better than serial processing models do, indicating that the nature of human information processing is intimately connected with the structure of the brain that produces it.

I have discovered some common threads in other people's attempts to integrate the vast sums of knowledge in their fields into theories of consciousness: everyone seems to agree that the brain uses parallel processing to produce consciousness, and that at the same time our experience of our own mental processes seems to indicate that consciousness is serial. Episodic experience is both global in its integration of all sensory modalities and the output of several higher-order processors, and unified because of its inability to think about many things at once. Sensory processing, on the other hand, can handle vast quantities of information in many modalities simultaneously. So at the same time many things are happening at once but they are experienced one at a time. Large amounts of information can thus be used to inform any particular conscious event. This ability to use multiple informational sources on single issues is the great strength of the brain, what allows it to produce intelligent, complex, conscious behavior.

What is consciousness? Consciousness is the various processes occurring in the brain over time, and nothing else. It is not a magical substance or property, it is not an unexplained new physical property, it is reducible to the physical properties

and processes of the brain, but not explainable wholly in those terms. The brain works by multiple roughly simultaneous content-fixations occuring in its massively parallel architecture. This structure enables it to deal with large amounts of information at the same time. The contents thus descriminated certainly feel like a unified entity because over and above the multiple parallel processing there is a serial process which deals with only a few things at a time. Descrimination of contents in any given domain occurs via a pandemonium-like process, in which those parts which "shout the loudest" (by virtue of experience, built-in strength, recency of activation, or additive stimulus from elsewhere in the brain) determine what content is discriminated. The serial nature of consciousness is produced by what might be called "probes" originating either in the environment or within the brain ("Now what do I do? Now what do I think about? What's for dinner?") which activate relevant content-fixers, often more than one at a time. Individual content-fixers in the brain may at any time contribute to the stream of conscious experience, and their participation in this stream is also determined by a pandemonium-type process. In humans, this serial process is closely tied to language and association. The combination of linguistic content-fixations and sensory content-fixations (including all the affectual components tied to them) creates a narrative-like stream of consciousness.

What mysteries remain? The serial processing component of consciousness is far less well understood than the smaller, more easily conceivable parallel processes which inform it. How discriminations about the world and about our internal processes are ordered and controlled is difficult to imagine, though pandemonium comes close. The fact that language plays such a large role in our self-definitions and in our conscious thought seems incredibly important, but has yet to be mechanistically explained. What is the role of language in consciousness? Can there be consciousness (in the human sense) without some form of language? Are the bits of developed concepts we acquire from culture really crucial to consciousness?

We can grasp how mechanistic sensory processes become phenomenal experiences. We can understand how parallel architectures hold information in a distributed fashion, and how the competition between feature detectors can discriminate the inchoate data of the world. We can rid ourselves of the misleading intuitions caused by the illusions of the Cartesian Theater and the Central Meaner. Eventually, we can understand consciousness.

The next chapter presents an experiment meant to further develop our knowledge of the functioning of the brain. This experiment approaches perception at the cognitive level, exploring the limits and parameters of human perception.

# Chapter 5

# Orwell vs. Stalin

One of the main points of Dennett's Multiple Drafts theory is that there can be no distinguishing between separate simultaneously active drafts of an experience until a narrative is precipitated by a probe. There is no fact of the matter as to which one is the "real" experience, as they are all candidates until one of them "wins" and makes it into the narrative. Furthermore, the timing of the probe has an effect on which one "wins," so in principle none of them is the "real experience." To ask that is to ask the wrong question, a question based on the assumption of a Cartesian Theater. The contrasting ideas, called Orwellian and Stalinesque, say that there is a fact of the matter. The Stalinesque theory says that your brain uses before-the-event information to alter your perception of the event, before the event becomes conscious. It is called Stalinesque because you are experiencing a kind of show trial: the outcome is determined by what has come before. The Orwellian theory, named after George Orwell, says that the evidence is doctored after its initial fleeting conscious experience—history is rewritten, so to speak, by your brain's Orwellian editor and so is misremembered. This distinction is important in his discussion of how memories are formed and influenced and is a crucial part of the Multiple Drafts theory.

Ned Block, in his review of *Consciousness Explained* ([Block, 1993], p. 187) calls

this idea "the argumentative heart of the book." George Mandler, in his review of Dennett's book, also believes the Orwellian/Stalinesque distinction is a major issue for Dennett, and he goes on to criticize Dennett's interpretation:

> The blurring of this distinction is achieved by a "trick," [which] involves the invocation of a very swift "wiping out" of some of the experiences that could construct particular memories. As a result of such "wiping out" Dennett asserts that there is no evidence available whether the editing occurred pre- or post-experientially. There is precious little support for such a "wiping out" process, and we do have available methods of unconscious priming and probing that might resolve whether these memories are as totally wiped out as Dennett would wish. ([Mandler, 1993], pp. 335-336)

Taking Mandler's interpretation and conclusions to heart, I asked him (via electronic mail) if he had some experiment in mind that utilized these "methods of unconscious priming and probing" and could test Dennett's assertions. He provided the basic ideas for the experiment I performed, an experiment that uses subliminal stimuli to attempt to alter people's judgement of supraliminal stimuli, and is meant to test whether pre-experiential or post-experiential stimuli have a greater effect.

After performing the experiment and then re-reading Dennett's book [Dennett, 1991] I realized that the experiment was not necessarily testing Dennett's theory, for Dennett could argue like this:

> Mandler is missing the point! The various content-fixations are not "experiences that could be used to construct particular memories." They are not experiences, there are no experiences, until a draft has been probed, until a narrative has been precipitated, until one interpretation "wins." What happens to the others? The respective brain areas move

> on to new problems, or are enlisted by the winning processing area, or some such. They don't lay down memories, they don't play a part in experience, their proprietary drafts have no further effects. They are not "wiped out" from memory, as memory is what is created by the "winning" draft.

I believe this kind of response is potentially damning to our original experimental hypothesis, and would make an interpretation based on it weak if not impossible. However, this experiment does test a potentially very important part of human perception at limited time scales: When people are forced to make swift judgements, what factors influence their perception? An answer to this question would be important for people like fighter pilots, who have to make snap judgements of order and identity based on limited information, information potentially presented subliminally. If their judgements can be affected differentially by stimuli which precede as opposed to stimuli which follow the supraliminal stimulus they react to, some redesign of instruments or policies for action may be called for. Furthermore, this experiment has relevance for questions about whether subliminal effects can have any effect at all, and if they do have an effect, whether subliminal stimuli presented before or after another (supraliminal) stimulus have a stronger effect. While this experiment does not test the "argumentative heart" of Dennett's book, as I had hoped, nonetheless it may help explain yet another part of the perceptual machinery that forms our phenomenological world.

The literature on subliminal effects is somewhat contradictory, and appears to depend heavily on methodological details. Duration of stimulus presentation; duration, type and use of post stimulus masks; type of test; and practice effects can all vary widely, and all have some effect on final results. An experiment by Daniel Orzech [Orzech, 1983] used tachistoscopic presentation of geometric figures (for 3msec each) to test for an exposure effect on preference and recognition. "Seeing" the geometric

figures, subliminally, was supposed to have created a preference for them as opposed to new figures, but was not supposed to have enabled the subjects to discriminate previously "seen" figures from new figures. His experiment failed to find an effect, contrary to the results of a similar experiment by Kunst-Wilson and Zajonc [Kunst-Wilson and Zajonc, 1980], although he followed their methods quite closely. The effect on preference should have been easier to find as it was based on an implicit test rather than the explicit test for recognition. Greenwald and Klinger [Greenwald and Klinger, 1990] tested whether subliminally presented and masked words would be more easily detected than similarly presented non-words, thus giving evidence for unconscious semantic discrimination. They also found no effect, contrary to the results of an experiment by Doyle and Leach [Doyle and Leach, 1988] which used a slightly different masking technique.

Positive results have also been found in semantic priming studies and in emotional-stimulus detection studies. Marcel [Marcel, 1983] found that subliminal masked priming words presented for 10 msec made detection of related words faster, e.g. "bank" primed detection of the words "river" and "money." A study by Làdavas et al. [Làdavas et al., 1993] found that subliminal masked stimuli that were either emotional (sexual or disgusting) or neutral presented for 20msec could be identified as emotional or neutral without conscious identification. As far as I have seen, however, the effects of arbitrary subliminal stimuli on color judgement have not been tested, nor have differential effects of pre-item versus post-item subliminal stimuli. Although a distinction between conscious and unconscious perception may be impossible in principle, as Dennett argues, the effects of stimuli which don't make their way into the narrative of episodic memory still seem to play a part in cognition, and exploration of this domain is important.

# 5.1 Method

## 5.1.1 Subjects

The subjects in this experiment were 51 Reed College undergraduates. Data on age, being relatively irrelevant, were not recorded. Students who needed corrective lenses used them. Subjects were recruited from the general student population immediately before the time of the session, told the experiment would only take 10-15 minutes, and were rewarded with tickets for a special lottery for participants in senior psychology experiments, as well as with candy (for those who desired it). Two subjects were dropped for some post hoc data analysis because they purposely responded the same way to every stimulus.

## 5.1.2 Apparatus

The stimuli were presented with a Macintosh IIvx computer, on a 13" Apple color monitor. The program for presenting the stimuli was PsyScope 1.0f8 from the psychology department at Carnegie Mellon University, a flexible tachistoscopic presentation program which allows exact timing of stimuli. All trials were taken in a soundproof room containing the computer, two chairs, and two tables. Responses were entered via the standard keyboard with no modifications. Subjects could ask questions of the experimenter during the session by speaking through the intercom system.

## 5.1.3 Stimuli

The stimuli were of three types, all presented in the center of the computer screen: 1) subliminal word stimuli, 2) masking stimuli, and 3) shape stimuli. The subliminal word stimuli (the words "green" and "purple") were in black, 12 point Palatino text and were presented for 10 msec. The masking stimuli, which immediately followed

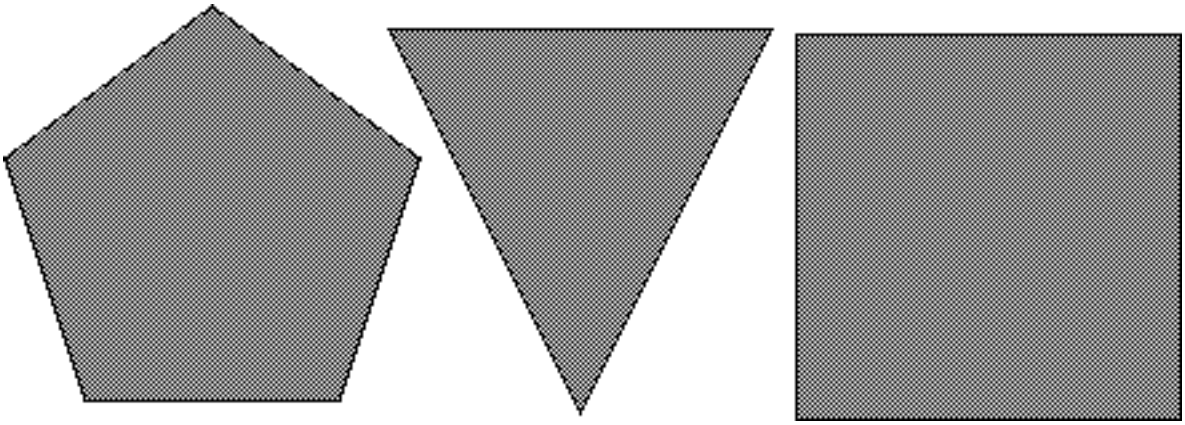

Figure 5.1: Examples of Shape Stimuli

the words, were six X's in 14 point bold shadowed text (**XXXXXX** ), presented for 20 msec. The shape stimuli were two isosceles triangles, point up and point down, a square, a pentagon, a hexagon and an octagon (see Figure 5.1). All shapes were approximately 2" by 2". These were drawn with Claris Macpaint and Macdraw and colored 50very fine-grained checkerboard of black and white), then saved as PICT files for import into PsyScope. The shapes were presented for 500 msec also in the center of the screen.

### 5.1.4 Procedure

Subjects were taken into the lab one at a time, and asked to fill out a consent form (see appendix A). After signing the consent form and filling out a lottery ticket, subjects were shown into the soundproof booth and seated themselves approximately 65 cm. from the screen. The instructions were presented on the computer screen:

> Welcome to my experiment!
>
> Please read this page, just so you know what to expect and how to respond. You will see some geometric figures flashed on the screen in

> front of you, along with some other stuff. These figures may be slightly tinged either green or purple, or you may see no color at all. Any tints will be very subtle, almost undetectable, so just respond whichever way feels most right. (By the way, the particular shapes have no relation to the colors, so you might see a green square one time and a purple square later.) In any case, after you see the shape please press either "g" for green or "p" for purple.[1]
>
> There will be 200 presentations, but each of them is very quick so this shouldn't take too long. If your eyes get tired, feel free to look away from the screen for a minute—just remember whether the last shape was green or purple. The program won't go on to the next trial until you press the "g" or "p" key, so if you just remember what you are going to press then you can rest your eyes for awhile—reaction time is not a factor.
>
> Thank you very much for participating in this experiment. A brief description of the theory and structure of this experiment will follow after all the trials, for the curious. Click the mouse button to begin.

and the subjects were told that they could ask questions through the intercom. Clicking the mouse button started the first of 200 trials, after a delay of 1000 msec.

The three kinds of stimuli (subliminal words, masks, and shapes) were arranged in three different ways to produce three trial types: Trial type 1: "pre-shape," n=50. Either the word "green" or the word "purple" was presented (randomly), followed by the mask, followed in turn by one of the six shapes (also selected randomly). Those particular color names were chosen because they are of similar length and both have

[1] In the "Levo" half of the second experiment, this phrase read "PLEASE PRESS EITHER "n" FOR GREEN OR "b" FOR PURPLE." In the "Dexter" half of the second experiment, this phrase read "PLEASE PRESS EITHER "b" FOR GREEN OR "n" FOR PURPLE."

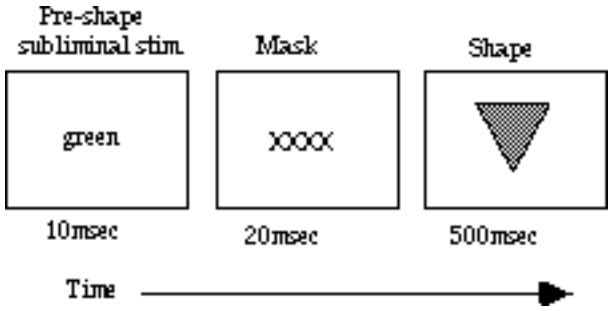


Figure 5.2: Pre-Shape Trials

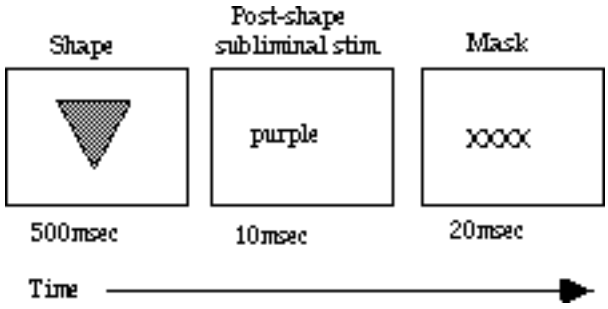


Figure 5.3: Post-Shape Trials

a descender as the first letter, thus making them harder to distinguish. Colors were chosen because they could be entirely arbitrary for the judgement, thus making a stronger case if any effects were found. Figure 5.2 is a schematic representation of these types of trials.

Trial type 2: "post-shape," n=50. These trials were similar to the pre-shape trials except for the order of stimuli. First the shape was presented, then the subliminal word, then the mask. Figure 5.3 is a schematic representation of these types of trials.

Trial type 3: Pre- and post-shape ("both"), n=100. In these trials the subliminal word stimuli and masks were presented both before the shape and after the shape. The selection of the word that came before the shape was independent of the selection of the word that came after the shape, and both selections were made

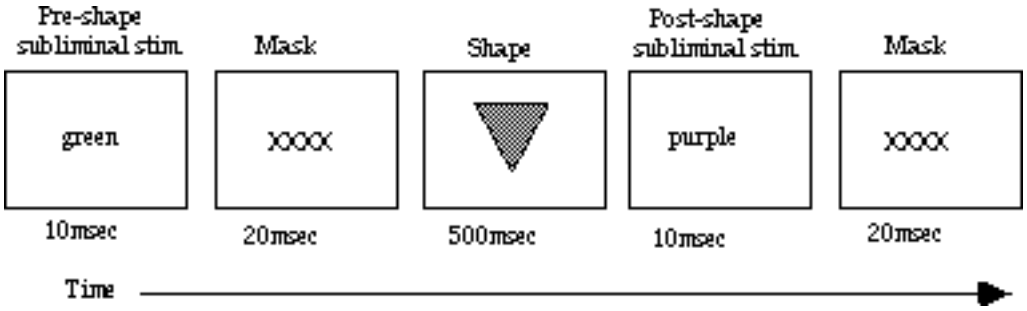


Figure 5.4: "Both" Trials

randomly. In other words, for all the trial types the two words ("green" and "purple") and the six shapes were randomized to produce all the possible combinations of stimuli. Figure 5.4 is a schematic representation of the "both" type of trial.

The trials were divided into three blocks, 50 each of the pre-shape and post-shape blocks and 100 of the "both" block, so that there would be an equal number of responses in each cell[2] for later analysis. The blocks were presented in random order, i.e. any one of the blocks could have come first for any subject. There were no breaks between them to indicate when one had ended and the next began. After all trials were complete, the debriefing form (see appendix B) was displayed on the screen. This debriefing form was meant to inform the subjects that they had been decieved about the coloring of the shapes and to explain to them the purpose of the experiment and the distinction between "Orwellian" and "Stalinesque" post-experiential memory revision.

I conducted two experiments differing only in which keys on the computer keyboard the subjects pressed to indicate their responses ("p" or "g" in the first experiment, "b" or "n" counterbalanced as to which one means green and which one means purple in the second experiment). This change was statistically insignificant (see results section), so I will collapse data from both experiments in the results.

[2]See table 5.1.4

Table 5.1: Experimental Design

| #1:Pre | **Resp.** | **Resp.** |
|---|---|---|
| **Stim.** | p | g |
| purple | **A** | **B** |
| green | **C** | **D** |

| #2:Pre | **Resp.** | **Resp.** |
|---|---|---|
| **Stim.** | p | g |
| purple | **E** | **F** |
| green | **G** | **H** |

| #3:Both | **Resp.** | **Resp.** |
|---|---|---|
| **Stim.** | p | g |
| pp | **S** | **T** |
| pg | **U** | **V** |
| gp | **W** | **X** |
| gg | **Y** | **Z** |

In summary, the design consisted of three different sets of trials with all possible combinations of stimuli producing two possible responses. Table 5.1.4 shows the format of the trials, where each possible combination of stimuli is a cell and each cell is assigned a letter for future reference. The pre-shape and post-shape groups were included to test for the effects alone rather than in competition: a difference in effectiveness of the pre- or post-shape words (i.e. between cells A+D and E+H in table 5.1.4). The interesting test in the 4x2 trials was in the pg and gp factors (cells U+V and W+X): which would have more influence, the word that came before the shape or the word that came after? The pp and gg factors were included as controls and as further tests for whether there was any subliminal influence at all. In addition, the results from the pre-shape trials and the post-shape trials could be compared with the "both" trials.

## 5.2 Results

Two of the blocks of trials contained two stimuli and two responses and one (the "both" trials) contained four groups of stimuli and two responses. The results from the 2x2 trials could be collapsed and analyzed with a chi-square procedure (and yield odds ratios) if there were no significant between-subject differences, and the 4x2 could be analyzed via ANOVA. Unfortunately there was a significant between-subjects interaction, so an alternate method of analysis was used, the general linear model.

The results of this experiment were analyzed in three stages. A form of the general linear model called the logistic regression model was used to find the effects of predictor variables on the subjects' response counts. This analysis was performed using the statistical package S+, and run on a DEC 5000. An autocorrelation for response lags was used to find whether individual subjects varied their responses randomly or with a pattern. This analysis was performed using a customized True-Basic program written by Alec Rogers and run on a Macintosh IIvx. Finally, odds ratios for the cell counts were computed, even though the logistic regression model showed them not to be significant. The odds ratios were analyzed with S+ as well.

### 5.2.1 Linear Regression Analysis

The linear regression model was used because the responses were binomial (either "p" or "g") and were compared with the cell counts and other factors in the experimental design. This model is appropriate because it does not require transformation or "bending" of the data to fit a linear model, and it doesn't assume a constant variance. Instead it uses reparametrization to induce linearity and allows nonconstant variance. Generalized linear models require two functions: a link function that describes how the mean depends on linear predictors, and a variance function. The

logistic regression model uses the logit link (which guarantees that $\mu$ is in the interval [0,1]) and the binomial variance function.

What this model measures is how well various factors predict the linearized data: in this case, how the two response factors, eight stimulus factors (two each for the pre and post blocks, four for the "both" block), forty-nine subject factors, and three group factors (one for the first experiment and two for the second) and all their interactions predict the cell counts. From this procedure we obtain the full residual deviance, which should be equal to zero. By restricting which factors and interactions are included in the model we obtain residuals which can be compared to the full model to derive p-values. When a restriction is performed, the p-value indicates whether the restriction removed a major source of variance or an insignificant source of variance. If the p-value is $>.05$ we can keep the restricted model, because it indicates that the factor removed in the restriction did not affect how well the model fits. In this experiment, the hope was that all sources of variance (and their interactions) except for the stimulus and the response could be removed, i.e. they would all have p$>$ 78.05. Unfortunately, this was not the case.

The full model's residual deviance was not equal to zero because the design is not fully saturated: each subject was not in each of the group factors but instead was nested within the groups. Therefore the starting "full model" was actually a restriction on the full model. This is not important, however, because of the result we got from the next restriction. The next restriction we performed was removing the group factors from all interactions, which tests if the group factors make any difference for the prediction of the cell counts. Since the p-value for this restriction was .135, we may keep the restriction and move on to other sources of variance—the group factor is not a good predictor.

The next restriction we tried was removing the three-way interaction between responses, subjects and stimuli. This would have left us with three two-way in-

teractions, between subjects and responses, subjects and stimuli, and stimuli and responses. This restriction could not be retained (p=0.000) so we have to keep the three way interaction in the model.

Removing the subject by stimulus interaction was sucessful, p=1.0, so we could keep this restriction. In other words, removing the subject by stimulus interaction does not affect how well the model fits—it is not a good predictor.

At this point we decided to test the main experimental hypothesis: that the stimulus by response interaction is a good predictor for the cell counts. This was a successful restriction (p=.092) so we can remove this interaction as a source of variance. However, it was close, so we decided to perform the odds ratio analysis with the understanding that its results were not significant.

Further restrictions were performed to try to see what the main sources of variance were. All factors other than the response by subject and the response by subject by stimulus interactions can be successfully performed, leading to the conclusion that a model based only on interactions between subjects and responses would predict the data as well as any other.

### 5.2.2 Autocorrelation Analysis

This analysis tests whether a given subject's responses could be predicted on the basis of a pattern in their responses. Correlations between any two, three, four, five or six responses were examined. With this analysis we hope to find no correlations, indicating that some other factor was responsible for the pattern of their responses. High correlations would indicate that the patten of their responses is a good predictor for the data. A perfect correlation at lag 1, for instance, would indicate that the subject was responding p,g,p,g,p,g . . . lag 2 would be p,p,g,g,p,p,g,g . . . and so on. Low correlations would indicate that subjects were behaving somewhat randomly, although assuming that the baseline for human binomial responses is

probably not justified. There is a statistic available for determining what r-values would indicate randomness, $2 * \sqrt{\frac{1}{n}}$, where n = number of responses. (Priestley, 1981, p. 340) This yields .14 for 200 responses. These data show that a number of subjects had very high lag correlations for many lags, which could explain why the subject factor was such a large predictor in the first series of statistics. Since this was an entirely post-hoc analysis and the data would have to be analyzed further before they could become useful (e.g. as another factor in the linear regression model), I did not use these correlations for anything but explanation of the between subjects differences. Appendix C contains the tables of autocorrelations for each lag (1-6) for each subject.

### 5.2.3 Odds Ratio Analysis

This analysis was performed in spite of the non-significant result of response by stimulus interaction obtained through the linear regression analysis. High odds ratios would mean that subjects were responding accurately to the stimuli, and a difference between the odds ratios for the different stimulus conditions would indicate whether the pre-shape or post-shape stimulus was more influential. In other words, the odds ratios tell us the probability that the subject will respond "g" to the stimulus "green" or "p" to the stimulus "purple." The odds ratios computed here represent the products of cells[3] A+D over B+C for the pre trials (A*D/B*C), cells E+H over cells G+F for the post trials, cells U+X over cells W+V for the "bothdiff" trials (those stimulus conditions where the subliminal stimulus that came before the shape was different from the one that came after), and cells S+Z over cells T+Y for the "bothsame" trials (those stimulus conditions where the subliminal stimulus that came before the shape was the same as the one that came after). Since the p-value from the linear regression analysis was <.10, some trend or tendency might be postulated.

[3]See table 5.1.4

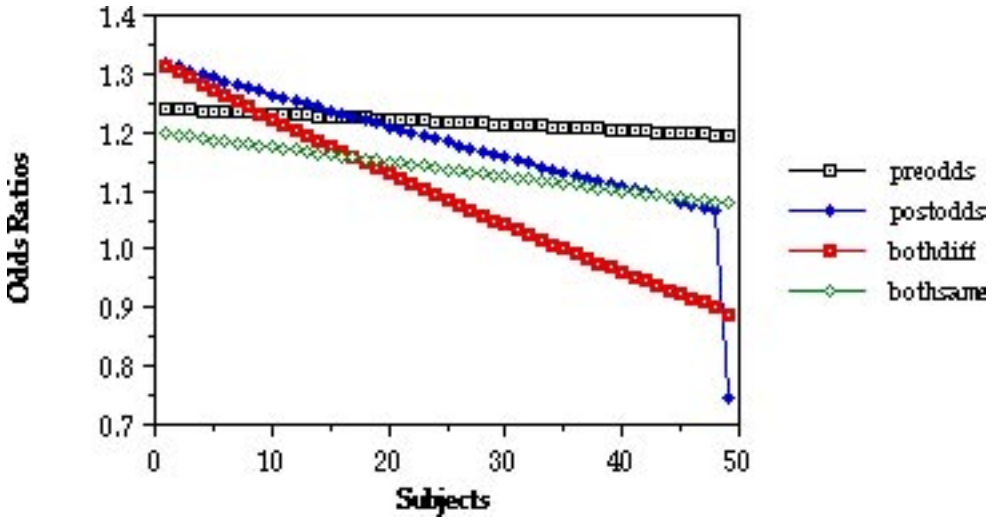


Figure 5.5: Sorted Odds Ratios By Subject and Stimulus Group

To find out which direction this trend might be in, I merely added the odds ratio values across subjects to obtain means. The means for these odds ratios are as follows:

| | |
|---|---|
| For the pre trials: | 1.217692 |
| For the post trials: | 1.178864 |
| For the bothdiff trials: | 1.090677 |
| For the bothsame trials: | 1.160779 |

All of these values are larger than 1.0 (even counts in all cells) but not by much. That is, for the pre trials there are 1.2 to 1 odds that the subject will respond "correctly" to the stimulus.[4] That the p-value for this interaction (stimulus by response) was as low as it was seems surprising, but is explained by the large amounts of data points collected, giving many degrees of freedom. Figure 5.5 is a plot of all of the individual odds ratios in sorted order, with the four different lines representing the four different odds ratios collected per subject.

[4]Not odds I would bet on!

## 5.3 Discussion

This experiment failed to find a significant effect of subliminally-presented words on subjects' color judgements. It did point to a trend toward a response by stimulus interaction, but the analysis of that interaction was also fairly inconclusive. Why did it fail? There are as many possible reasons as there were factors in the experiment (subjects, responses, stimuli, groups) but the linear regression analysis makes it fairly clear that the subject by response factors contained the main reason. Since the responses were binomial, the keys pressed varied across groups, and the group factors were not significant predictors, it seems likely that the variability from subject factors was the reason for the nonsignificant effect. Of course, if the effect had been strong it would have overshadowed the between subjects variance, so the stimuli and their presentation may also be to blame.

Individual subjects have very different perceptual faculties, so the ability of some subjects to unconsciously perceive the stimuli and the inability of others would tend to cancel each other out. Post-hoc removal of those subjects whose responses failed to meet some criterion seems like changing the data to fit the hypothesis, but some case could be made for this kind of move in the interests of discovering whether there was some directional effect between the pre-shape and post-shape blocks. Figure 5.5 shows, however, that even the subjects whose odds ratios were highest did not respond reliably to the stimuli: odds of 1.3 to 1 are not odds worth betting on.

Individual subjects also may have reacted in different ways to the task itself. Some of the subjects mentioned their reactions. The fact that the shapes themselves were in fact just grey, with no tinge of color at all, may have been frustrating, causing subjects to behave perversely; i.e. responding to many trials in a row with the same judgement ("They're all grey anyway, so I'll just press "p" for a while"). Subjects could also respond to this task either randomly or with high lags. Even

those subjects who decided that there was no way to discriminate between the two "colors" of grey and so decided to just "go with their feelings" didn't yield high odds ratios, however.

Yet another subject variable that could have affected the responses was some kind of judgement about the color grey right at the start. Does grey look more like green or like purple? Of course it is actually a neutral color, but people are not wavelength-discriminators, and whichever way they would normally answer this question could have affected their responses more than the stimuli.

The stimuli themselves were probably a cause of failure since their effect was not large enough to overcome between-subject differences. Was the presentation time too long? It was within the normal range of times in the literature, and even at 10 msec some subjects reported seeing the words once or twice. Was the presentation time too short? This seems quite possible, especially given the double-masking effects of the mask and the shape, except that if the odds ratios say anything they say that the pre-shape trials had the strongest effect.

One potential problem with the overall design may have been the very short interstimulus intervals. As soon as the subjects responded to one trial, the next one began. Of course, they were told that they could wait as long as they liked before responding, but subjects in general responded immediately after the end of the trials. It could be that the subliminal stimulus did not have time to affect judgement, that the content-discrimination (based on very hard to see information) by the combined visual and linguistic processing systems did not happen in time to affect the response. Reaction time data might provide some way of distinguishing between these hypotheses. (RTs were recorded but the timing was found to be inaccurate, so they were not used. They also would have further complicated the data analysis.)

What positive statements can be made about the results of this experiment?

The odds ratios are very slightly in favor of the stimulus-response matches over mismatches. The differences in odds ratios between the pre-shape and post-shape blocks and the odds ratio for the "bothdiff" cells in the "both" block also indicate a slightly stronger pre-shape effect. This could be explained by determining the teleological function of sensory processing: What is it for? If the senses evolved to accomplish the function of "future-prediction" they would need to be able to make quick judgements about what was to come from whatever information they could extract from the environment. Thus subtle stimuli which indicate some important factor in the environment should be handled by fast, specific-purpose modular processors, and their descriminations of these stimuli should be linked to relevant information. Perhaps the use of completely arbitrary stimuli in this experiment made content discrimination too difficult because the stimuli don't activate any pre-existing detectors or connections from those detectors: there was no need for the kind of detector that would be able to use these kinds of stimuli to evolve.

### 5.3.1 Suggestions for Future Research

Given the failure of this experiment and the difficulty in assigning reasons for it several different avenues could be explored. One change could be in the saliency of the stimuli, by making them larger or presenting them for longer periods of time. If subjects were required to respond as quickly as possible to the stimuli, some effect of pre- versus post-supraliminal influence might come out: forced to respond quickly, and given conflicting messages, subjects might show a stronger effect. Lengthening the inter-stimulus interval is another possibility. If sufficient time is allowed for conflicting information to be discriminated, we might see a more reliable effect.

A more radical restructuring of this experiment could be performed to make it more "naturalistic": Subliminal stimuli could be used that have some relation to the supraliminal stimuli, and so perhaps take advantage of mechanisms that exist in

the brain for this kind of discrimination. The challenge for this type of experiment would be in finding environmentally relevant stimuli that can be cued or primed in opposite ways. Of course the experiment by Marcel (1983) did this type of priming with words, but he did not investigate the pre- versus post-stimulus effect, and words seem a little artificial.

Finally, I would like to suggest some research that would test ideas presented in the main body of this thesis. Dennett believes that human consciousness depends heavily on languages and the memes transmitted through language. This seems to indicate that people who don't speak a language aren't conscious in the same way that normal humans are, that their consciousness is fundamentally different, perhaps more like the consciousness of a chimpanzee! An experiment to test this would be difficult to arrange because almost everyone in the world speaks a language and participates in a culture, but it might be possible to find deaf people who have not learned a sign language or written language and test them in several ways. Do they have a self-concept? Can they use logical (or heuristical) methods to solve problems? Can they make long-term plans for themselves? Are they easily distracted or incapable of focus?

# Appendix A: Consent Form

**Consent Form:**

I agree to participate in this experiment on human perception. I understand that no harmful stimuli will be used. I understand that the results and data from this experiment will be kept completely confidential, and that I may quit at any time.

If I have any questions about this experiment I understand that I may ask the experimenter anything at all. Any concerns about this experiment may be addressed to Professor Enriqueta Canseco-Gonsalez in the Psychology building (x425) or Professor Albyn Jones, chair of the Human Subjects Research Review Committee.

_____ I have read and understand this form.
(Signature of participant)

_____ I have presented this form.
(Signature of experimenter)

# Appendix B: Debriefing Form

**Debriefing Form:**

Thanks for sitting through all that! And now, a few words about this experiment: I am trying to distinguish between two different theories about how your brain's perceptual processes and memory/consciousness processes work.

One theory, called the Stalinesque theory, says that your brain uses before-the-event information to alter your perception of the event. It is called Stalinesque because you are experiencing a kind of show trial: the outcome is determined by what has come before.

The other theory is called Orwellian, after George Orwell. The idea here is that the evidence is doctored after the experience—history is rewritten, so to speak, by your brain pre-conscious processes before you become conscious of them.

To test between these, I showed you shapes and said that they were possibly tinted one color or another, as well as the words "green" and "purple" followed by a mask (XXXXXX). In actual fact the shapes were all just grey, but I wanted to see if the color words could make you think they were tinted. There were three groups of trials: One where the words were presented before the shapes, one where they were presented after the shapes, and one group where the words came both before and after the shapes. By analyzing what you said about the shapes' "colors" I hope to find evidence for or against one of these theories. Once again, thank you very much, and feel free to ask me any questions. Oh yeah—please don't tell your friends about this, because those words are supposed to be subliminal, and if they know... Click the mouse to end.

# Appendix C: Autocorrelation Tables

Autocorrelations by Subject (Experiment 1)

| Subject | Lag 1 | Lag 2 | Lag 3 | Lag 4 | Lag 5 | Lag 6 |
|---|---|---|---|---|---|---|
| 1 | -.0213 | .2293 | .0576 | .1823 | .0174 | .0574 |
| 2 | .6851 | .6040 | .6118 | .6096 | .5871 | .4833 |
| 3 | .0950 | .1295 | .0157 | -.0286 | -.0238 | -.1197 |
| 4 | -.3685 | .2576 | -.1741 | .0301 | -.0251 | .1010 |
| 5 | .1065 | .0098 | -.1624 | .1312 | .1290 | -.0602 |
| 6 | -.0915 | .0118 | -.0825 | -.0486 | -.1241 | .0823 |
| 7 | -.0327 | .0255 | .1836 | -.1015 | .1120 | -.0505 |
| 8 | .0010 | .0764 | .0526 | -.0624 | .0633 | .1591 |
| 9 | .0461 | -.0128 | -.0787 | -.0791 | .1925 | .0574 |
| 10 | .3614 | .2840 | .3542 | .3187 | .0979 | .1250 |
| 11 | .4083 | .1377 | .0638 | .2091 | .1843 | -.0031 |
| 12 | .4015 | .2563 | .2308 | .1810 | .1794 | .2503 |
| 13 | .0498 | -.0512 | -.0465 | .1005 | .0278 | -.0681 |
| 14 | -.2844 | -.0346 | .0494 | -.0250 | -.0201 | -.0554 |
| 15 | .2318 | .1984 | .1448 | -.0294 | .0160 | .0710 |
| 16 | -.0533 | -.1019 | .1050 | -.0063 | .1164 | .0672 |
| 17 | -.2003 | .0503 | -.1435 | .1101 | -.0854 | .0406 |
| 18 | -.0537 | .1239 | -.0228 | -.0836 | -.0129 | -.0300 |
| 19 | .0028 | -.0732 | -.0280 | -.1050 | .0229 | -.0219 |
| 20 | .1837 | .0905 | .1210 | .0762 | .1147 | -.0280 |
| 21 | -.3142 | .2017 | -.1195 | .0641 | -.0707 | .1249 |
| 22 | -.0055 | -.0209 | -.0973 | -.0316 | -.1829 | -.0462 |
| 23 | .1156 | .2723 | .2324 | .1135 | .1602 | .1273 |

Autocorrelations by Subject (Experiment 1)

| Subject | Lag 1 | Lag 2 | Lag 3 | Lag 4 | Lag 5 | Lag 6 |
|---|---|---|---|---|---|---|
| 24 | -.0166 | .0709 | .0845 | .0064 | .1082 | .0800 |
| 25 | .2121 | .0795 | .0253 | .1011 | .0771 | .0931 |
| 26 | -.0947 | .5141 | .0206 | .5069 | .0313 | .4228 |
| 27 | -.1962 | .1527 | -.1585 | .0049 | .0400 | .0253 |
| 28 | .3630 | .4251 | .1970 | .1947 | .0200 | .0684 |
| 29 | .0000 | .0935 | -.0102 | -.0848 | .1001 | .0858 |

Autocorrelations by Subject (Experiment 2)

| Subject | Lag 1 | Lag 2 | Lag 3 | Lag 4 | Lag 5 | Lag 6 |
|---|---|---|---|---|---|---|
| 1 | .8248 | .7305 | .6561 | .6438 | .5796 | .5149 |
| 2 | .5486 | .4725 | .2918 | .2144 | .1444 | .2605 |
| 3 | .3616 | -.1057 | -.1514 | -.1073 | -.0225 | .0524 |
| 4 | .1973 | -.1156 | -.2828 | -.1900 | -.0327 | -.0073 |
| 5 | -.0699 | .0231 | .0481 | -.0967 | -.0321 | -.0474 |
| 6 | -.0410 | -.1461 | -.0603 | .0060 | -.0303 | .0488 |
| 7 | .3423 | .3291 | .0941 | .0593 | .0649 | .0817 |
| 8 | .5304 | .3674 | .2366 | .1380 | .2349 | .1361 |
| 9 | .0566 | -.0171 | .0476 | -.0069 | .1784 | -.0067 |
| 10 | -.0234 | -.1026 | -.0138 | .0416 | .0602 | -.1570 |
| 11 | .1061 | -.0765 | -.2017 | -.1377 | .0567 | .0823 |
| 12 | .1757 | .0416 | .0468 | .2832 | .1678 | .2257 |
| 13 | .1343 | -.0184 | -.0732 | .0607 | .0660 | .0512 |
| 14 | -.1484 | .0728 | -.2590 | .1934 | -.1610 | .1450 |
| 15 | .3701 | .2456 | .3751 | .2288 | .2468 | .2950 |
| 16 | .0490 | -.0838 | -.1683 | .0349 | -.0900 | .0352 |
| 17 | -.6662 | .4746 | -.3243 | .3101 | -.2153 | .3031 |
| 18 | -.0653 | .0905 | -.1567 | .0812 | .0346 | .0309 |
| 19 | .1658 | .2306 | .1207 | .1043 | .1107 | .1173 |
| 20 | .7456 | .6827 | .6180 | .5770 | .6339 | .6089 |

# Bibliography

[Baddeley and Wilson, 1985] Baddeley, A. and Wilson, B. (1985). Phonological coding and short-term memory in patients without speech. *Journal of Memory and Language*, 24:490–502.

[Bahrick, 1984] Bahrick, H. P. (1984). Semantic memory content in permastore: Fifty years of memory for spanish learned in school. *Journal of Experimental Psychology: General*, 113:1–29.

[Bahrick et al., 1975] Bahrick, H. P., Bahrick, P. O., and Wittlinger, R. P. (1975). Fifty years of memory for names and faces: A cross-sectional approach. *Journal of Experimental Psychology: General*, 104:54–75.

[Block, 1980] Block, N., editor (1980). *Readings in Philosophy of Psychology: Volume One*. Harvard University Press, Cambridge.

[Block, 1993] Block, N. (1993). Review of consciousness explained. In [Dennett, 1991], pages 181–193.

[Boyd et al., 1991] Boyd, R., Gasper, P., and Trout, J. D., editors (1991). *The Philosophy of Science*. MIT Press, Cambridge, MA.

[Braitenberg, 1984] Braitenberg, V. (1984). *Vehicles: Experiments in Synthetic Psychology*. MIT Press, Cambridge, MA.

[Campbell, 1970] Campbell, K. (1970). *Body and Mind*. University of Notre Dame Press, Notre Dame.

[Carlson, 1991] Carlson, N. R. (1991). *Physiology of Behavior, 4th ed.* Allyn and Bacon (a Division of Simon and Shuster, Inc.), Needham Heights.

[Churchland, 1981] Churchland, P. M. (1981). *Eliminative materialism and the propositional attitudes*, pages 206–223. In [Lycan, 1990].

[Churchland, 1993] Churchland, P. S. (1993). *Neurophilosophy*. MIT Press, Cambridge, MA.

[Conway et al., 1991] Conway, M. A., Cohen, G., and Stanhope, N. (1991). On the very long-term retention of knowledge acquired through formal education: Twelve years of cognitive psychology. *Journal of Experimental Psychology: General*, 120, 4:395–409.

[Crick and Koch, 1990] Crick, F. and Koch, C. (1990). Towards a neurobiological theory of consciousness. *Seminars in the Neurosciences*, 2:263–275.

[Davidson et al., 1986] Davidson, R. J., Schwartz, G. E., and Shapiro, D., editors (1986). *Consciousness and Self-regulation, vol. 4*. Plenum Press, New York.

[Davies and Humphreys, 1993] Davies, M. and Humphreys, G. W., editors (1993). *Consciousness*. Blackwell, Oxford, UK.

[Dennett, 1991] Dennett, D. (1991). *Consciousness Explained*. Little, Brown & Co., Boston.

[Doyle and Leach, 1988] Doyle, J. R. and Leach, C. (1988). Word superiority in signal detection: Barely a glimpse yet reading nonetheless. *Cognitive Psychology*, 20:283–318.

[Fodor, 1983] Fodor, J. A. (1983). *The Modularity of Mind*. MIT Press, Cambridge, MA.

[Fodor, 1986] Fodor, J. A. (1986). *Banish DisContent*, pages 420–438. In [Lycan, 1990].

[Gleitman, 1986] Gleitman, H. (1986). *Psychology: Second Edition*. W. W. Norton & Co., New York.

[Greenwald and Klinger, 1990] Greenwald, A. G. and Klinger, M. R. (1990). Visual masking and unconscious processing: Differences between backward and simultaneous masking? *Memory & Cognition*, 18:430–435.

[Hofstadter and Dennett, 1981] Hofstadter, D. R. and Dennett, D. C. (1981). *The Minds I*. Bantam, New York.

[Jackson, 1982] Jackson, F. (1982). *Epiphenomenal qualia*, pages 469–477. In [Lycan, 1990].

[K. and Valenstein, 1985] K., H. and Valenstein, E., editors (1985). *Clinical Neuropsychology*. Oxford University Press, New York.

[Kunst-Wilson and Zajonc, 1980] Kunst-Wilson, W. R. and Zajonc, R. B. (1980). Affective discrimination of stimuli that cannot be recognized. *Science*, 207:557–558.

[Làdavas et al., 1993] Làdavas, E., Cimatti, D., Del Pesce, M., and G., T. (1993). Emotional evaluation with and without conscious stimulus identification: Evidence from a split-brain patient. *Cognition & Emotion*, 7:95–114.

[Levine, 1993] Levine, J. (1993). *On leaving out what it's like*, pages 121–136. In [Davies and Humphreys, 1993].

[Lycan, 1990] Lycan, W. G., editor (1990). *Mind & Cognition.* Basil Blackwood, Ltd., Oxford, UK.

[Mandler, 1993] Mandler, G. (1993). Review of consciousness explained. *Philosophical Psychology*, 6, 3:335–338.

[Marcel, 1983] Marcel, A. J. (1983). Conscious and unconscious perception: Experiments on visual masking and word recognition. *Cognitive Psychology*, 15:197–237.

[McCarthy and Warrington, 1990] McCarthy, R. and Warrington, E. (1990). *Cognitive Neuropsychology: A Clinical Introduction.* Academic Press, Inc., San Diego.

[McGinn, 1991] McGinn, C. (1991). *The Problem of Consciousness.* Basil Blackwood, Ltd., Oxford, UK.

[Medin and Ross, 1992] Medin, D. L. and Ross, B. H. (1992). *Cognitive Psychology.* Harcourt Brace Jovanovich College Publishers, New York.

[Nagel, 1974] Nagel, T. (1974). *What is it like to be a bat?*, pages 159–163. In [Block, 1980].

[Nagel, 1991] Nagel, T. (1991). What we have in mind when we say we're thinking. *The Wall Street Journal*, Nov. 17:A12.

[Neisser, 1984] Neisser, U. (1984). Interpreting harry bahricks discovery: What confers immunity against forgetting? *Journal of Experimental Psychology: General*, 113:32–35.

[Nieuwenhuys et al., 1988] Nieuwenhuys, R., Voogd, J., and van Huijzen, C. (1988). *The Human Central Nervous System.* Springer-Verlag, New York.

[Orzech, 1983] Orzech, D. (1983). This one or that?: The effects of exposure on stimulus preference and recognition. Master's thesis, Reed College.

[Pashler, 1993] Pashler, H. (1993). Doing two things at the same time. *American Scientist*, 81:48–55. includes refs and summaries of Pashler's 1984, 1989, 1991 papers.

[Penrose, 1989] Penrose, R. (1989). *The Emperors New Mind: Concerning Computers, Minds, and the Laws of Physics*. Oxford University Press, New York.

[Schwartz and Reisberg, 1991] Schwartz, B. and Reisberg, D. (1991). *Learning and Memory*. W. W. Norton and Company, New York.

[Schwartz, 1990] Schwartz, M. F., editor (1990). *Modular Deficits in Alzheimer-Type Dementia*. MIT Press, Cambridge, MA.

[Searle, 1984] Searle, J. (1984). *Minds, Brains and Science*. Harvard University Press, Cambridge, MA.

[Shalice, 1991] Shalice, T. (1991). Précis of *from neuropsychology to mental structure*. *Behavioral and Brain Sciences*, 14:429–469.

[Tversky and Kahneman, 1974] Tversky, A. and Kahneman, D. (1974). Judgement under uncertainty: Heuristics and biases. *Science*, 185:1124–1131.

[Tversky and Kahneman, 1981] Tversky, A. and Kahneman, D. (1981). The framing of decisions and the psychology of choice. *Science*, 211:453–458.