Experimental Knowledge in Cognitive Neuroscience: Evidence, Errors, and Inference

Mahir Emrah Aktunc

Dissertation submitted to the faculty of the Virginia Polytechnic Institute and State University in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

In

Science and Technology Studies

Deborah G. Mayo (chair)

Richard M. Burian

Aris Spanos

Lydia K. Patton

July 2, 2011

Blacksburg, VA

Keywords: Evidence, Inference, Cognitive Neuroscience, Severe Tests, Error Probabilities, Models of Inquiry, Experimental Knowledge

Experimental Knowledge in Cognitive Neuroscience: Evidence, Errors, and Inference

Mahir Emrah Aktunc

ABSTRACT

This is a work in the epistemology of functional neuroimaging (fNI) and it applies the error-statistical (ES) philosophy to inferential problems in fNI to formulate and address these problems. This gives us a clear, accurate, and more complete understanding of what we can learn from fNI and how we can learn it.

I review the works in the epistemology of fNI which I group into two categories; the first category consists of discussions of the theoretical significance of fNI findings and the second category discusses methodological difficulties of fNI. Both types of works have shortcomings; the first category has been too theory-centered in its approach and the second category has implicitly or explicitly adopted the assumption that methodological difficulties of fNI cannot be satisfactorily addressed. In this dissertation, I address these shortcomings and show how and what kind of experimental knowledge fNI can reliably produce which would be theoretically significant.

I take fMRI as a representative fNI procedure and discuss the history of its development. Two independent trajectories of research in physics and physiology eventually converge to give rise to fMRI. Thus, fMRI findings are laden in the theories of physics and physiology and I propose how this creates a kind of useful theory-ladenness which allows for the representation of and intervention in the constructs of cognitive neuroscience. Duhemian challenges and problems of underdetermination are often raised to argue that fNI is of little, if any, epistemic value for psychology. I show how the ES notions of severe tests and error probabilities can be applied in epistemological analyses of fMRI. The result is that hemodynamic hypotheses can be severely tested in fMRI experiments and I demonstrate how these hypotheses are theoretically significant and fuel the growth of experimental knowledge in cognitive neuroscience.

Throughout this dissertation, I put the emphasis on the experimental knowledge we obtain from fNI and argue that this is the fruitful approach that enables us to see how fNI can contribute to psychology. In doing so, I offer an error-statistical epistemology of fNI, which hopefully will be a significant contribution to the philosophy of psychology.

For

Hulki Aktunç

1949 – 2011

Poet, Writer, Essayist, Lexicologist,

and

Beloved Father

Acknowledgements

Throughout the years I worked on this dissertation, I have had the support of incredible people and it is my duty and pleasure to offer my sincerest feelings of gratitude for their support. First and foremost, I have to thank my advisor, Professor Deborah Mayo, without whose patience, wisdom, and trust in me, this work would have never been completed. I have had the privilege of having the best committee of advisors any graduate student can wish for. I thank Professor Richard Burian for his wise guidance in long and immensely fruitful discussions, philosophical and otherwise; Professor Aris Spanos for his great talent in making me understand highly complex philosophical and statistical notions; and Professor Lydia Patton for her meticulous appraisals of my chapters. The support of my committee has made me a better philosopher of science as well as a better person.

I am grateful to Professor Ellsworth Skip Fuhrman, the chair of the Department of Science and Technology in Society, for his invaluable support and friendship. I also have to thank Professor James Klagge, the chair of the Department of Philosophy, for giving me the opportunity to teach logic and philosophy, which has been immensely joyful for me over the years.

It was my privilege to be a member of the philosophical and STS community at Virginia Tech where I have studied with highly talented teachers and students. My interactions with all of them have been the most enriching experience. I thank Jean Miller for her undying support and friendship. I specially thank professors Joe Pitt, Gary Downey, and Matthew Goodrum not only for being incredible teachers but also for being wonderful friends.

Without the significant help of the administrative staff, Karen Snider and Crystal Harrell at STS and Terry Zapata and Leisa Osborne at Philosophy, my life would have definitely been much harder.

On a personal level, I am grateful for all the friends I have had in Blacksburg who have shared the graduate life with me and kept their friendship and support long after they moved on with their lives outside Blacksburg. My lifelong friend Özgür Gen has never been to Blacksburg, but he has supported me from afar and he is the best friend anyone can wish for.

Throughout my years in pursuance of this degree, my family had to make great sacrifices. My mother, Semra Aktunç, had to live with not seeing her beloved son for years at a time and there are no words in any language that I can describe my gratitude for the love, wisdom, and support of the greatest mother in the world. My father, Hulki Aktunç, had to be satisfied with long philosophical, historical, and literary conversations on the phone as this was our greatest joy as father and son. My brother, Uluğ Aktunç, missed me greatly as I missed him. I cannot describe my love and gratefulness for them and all other members of my family for being in my life.

The greatest happiness in this life is having a soul mate and I am one of those lucky people who have found one. There is no way that I can express my love for my soul mate, Esra Ağca, and my thankfulness for her being my ever strong anchor to all that is good in this life.

Table of Contents

Chapter One:
Voodoo Correlations, Salmon Thoughts, and the Promised Science of fMRI…….1

1.1: Introduction ……1
1.2: The Epistemology of Functional Neuroimaging ……6
1.2.1: Questions on The Theoretical Significance of fNI Findings…….....7
1.2.2: Discussions of the Methodological and Inferential Difficulties of fNI…...14
1.2.3: Some Shortcomings of The Philosophical Literature on fNI……..20
1.3: The Error-Statistical Account and Severe Tests……...22
1.4: Salmon Thoughts and Voodoo Correlations……..28
1.4.1: Salmon Thoughts…….29
1.4.2: Voodoo Correlations……...31
1.4.3: What Would the Error-Statistician Say?.....38

Chapter Two:
A Concise History of fMRI: Theory-Ladenness of a Useful Kind……...45

2.1: Introduction……45
2.2: The Beginnings of Cognitive Neuroscience……...46
2.3: History of fMRI……..49
2.3.1: The Discovery of Nuclear Magnetic Resonance…….49
2.3.2: Magnetic Resonance Imaging in Biology and Medicine…….53
2.3.3: Blood-Oxygenation-Level-Dependent Contrast and functional MRI…….55
2.4: Theory-Ladenness Of A Useful Kind…….60

Chapter Three:
Primary Models in fMRI: Determining What Is and What Is Not Underdetermined By Data…….71

3.1: Introduction……71
3.2: The Hierarchical Framework of Models of Inquiry and Severe Tests……...74
3.3: Primary Models in fMRI……82
3.3.1: Statistical Hypothesis Testing and Error Probabilities……85
3.3.2: Significance Tests in fMRI……..89
3.3.3: Recognition Memory in the Hippocampus……..95
3.4: The Theoretical Significance of fMRI Findings……...100

Chapter Four:
Experimental and Data Models in fMRI: Tackling Duhemian Problems……108

4.1: Introduction……...108
4.2: Experimental Models……112
4.2.1: Power of fMRI Scanners As A Source of Error……115
4.2.2: Dealing With Neuroanatomical Variability……...121
4.3: Data Models……..129
4.3.1: Preprocessing of fMRI Data……..129
4.3.2: Statistical Modeling of fMRI Data……....138
4.3.3: Multiple Testing and Thresholding in fMRI……145
4.4: Voodoo Correlations Revisited……151

Chapter Five:
Experimental Knowledge and Progress in Cognitive Neuroscience…….156

5.1: A Brief Recap……...157
5.2: Experimental Knowledge and Progress in Cognitive Neuroscience……162
5.2.1: Experimental Knowledge of Instruments and Procedures……162
5.2.2: Experimental Knowledge of Hemodynamic Substrates of Cognition…...165

Bibliography……...176

Chapter One:

Voodoo Correlations, Salmon Thoughts, and the Promised Science of fMRI

> *In the physical sciences, the usual result of an improvement in experimental design, instrumentation, or numerical mass of data, is to increase the difficulty of the "observational hurdle" which the physical theory of interest must successfully surmount; whereas, in psychology and some of the allied behavior sciences, the usual effect of such improvement in experimental precision is to provide an easier hurdle for the theory to surmount.* Hence what we would normally think of as improvements in our experimental methods tend (when predictions materialize) to yield *stronger* corroboration of the theory in physics, since to remain unrefuted the theory must have survived a more difficult test; by contrast, such experimental improvement in psychology typically results in a *weaker* corroboration of the theory, since it has now been required to survive a more lenient test.
>
> Paul E. Meehl, 1967 (pp.103-104)

1.1: Introduction

Questions relating to methods and inferential procedures in psychological science have occupied the minds of psychologists and philosophers for a long time. Foundational discussions, in the very beginnings of the establishment of psychology as a separate discipline, were centered on the methodology to be chosen for the new science. For example, one issue was whether or not introspection was an acceptable method which the behaviorists strongly rejected. This debate ended with the temporary victory of the methodological behaviorists, who dominated psychology until the "cognitive revolution" of the 1950's and 1960's, which adopted the experimental methodology of the

behaviorists, but at the same time *allowed* inferences from behavioral findings to certain conclusions about psychological entities and processes that presumably take place in the human mind/brain.

Along the way, several practicing psychologists have written about the methodological problems that arose in psychological research on humans; the work of Paul Meehl, who is quoted above, provides a good example of the worries and troubles a working psychologist had about the methods of the discipline. Meehl's central worry in the above quote is that as experimental methods and instruments of psychological science improve, it seems that it becomes easier for its substantive theories to be corroborated by experimental results. Meehl was concerned that on the basis of results of statistical significance tests psychologists would infer the truth of the whole of large scale theories of psychology, for example, Freudian theory. This is a kind of fallacy of rejection as defined by Deborah Mayo and Aris Spanos (2006). If an experimental result is statistically significant, the null hypothesis is rejected and the alternative hypothesis, which is entailed by some substantive theory, is accepted. The fallacy is inferring from this statistical result that one's substantive theory gains direct quantitative support confounding the statistical alternative hypothesis and the substantive theory. Of course, Meehl is quite right in stating that such easy-to-obtain corroborations do not provide strong evidence for substantive hypotheses. Meehl proposes a remedy for this problem that focuses on statistical significance tests: he suggests that psychologists should prefer point-prediction hypotheses instead of directional alternative hypotheses in significance tests. However, he also acknowledged that psychological science was just not as developed as physics to have substantive theories that could generate point-prediction

hypotheses, at least not in 1967 when he published the above article. Looking at the current state of affairs in methodological discussions in psychology, I have a hunch that had Meehl been alive today, he would retain this idea and perhaps even think that psychological science has gotten worse as it developed further to have at its disposal tools such as functional magnetic resonance imaging (fMRI), the use of which sometimes unfortunately leads to methodological controversies such as the "voodoo correlations" debate that I discuss later in this chapter. This is indeed a common theme in the history of psychological science; as with most sciences, researchers invent new and improved methods for data collection. Consequently, new ways of statistical analysis geared toward the types of data obtained from these new methods are devised. This kind of scientific development is usually accompanied by methodological debates and, in some cases, debates about the same issue resurface over decades in ostensibly new forms. Debates about statistical significance tests provide an informative picture: Meehl raised criticisms of the use of significance tests in his 1967 paper quoted above. After Meehl, questions on the use of significance tests prevailed but were not strongly emphasized until the 1990's when a group of staunch critics called for a ban on significance tests in psychology (Hagen, 1997; 1998; Krantz, 1999; Wilkinson & Task Force on Statistical Inference, 1999). This was a heated debate about crucial methodological questions for the conduct of psychological science. However, it has mostly faded away unresolved. The majority of psychologists still use significance tests and the misuse of statistical methods in psychology is no less problematic today than when a ban on these tests was discussed. If anything, it may be even more problematic as the scientific community, as well as the popular news media, have witnessed the controversy of "voodoo correlations," which

was centered around a provocative commentary by Edward Vul and his colleagues (2009) suggesting that a lot of the correlation coefficients calculated on neuroimaging data in social neuroscience were based on flawed analyses. Social neuroscience uses fMRI as its main experimental paradigm and methodological criticisms of the use of functional brain imaging tools, such as fMRI, in psychological science are not new.[1]

Van Orden and Paap's critique of functional neuroimaging (fNI) appeared in *Philosophy of Science* as early as 1997 when the very idea of using fNI tools in psychological science was still new and cognitive neuroscience was just beginning to come of age as a separate discipline within psychology and cognitive science. Yet, these authors were quick to dismiss fNI as a useful paradigm for research saying that for fNI techniques to work one first had to have a full blown and true modular theory of the mind and a series of assumptions that are probably false. William Uttal was another early critic, who built on the criticisms of Van Orden and Paap, and he initiated another heated debate with his book *The New Phrenology* (2001), which attracted considerable attention from philosophers of science. If we step back and look for a moment, we can see that Uttal and Vul and his colleagues can be thought of as contemporary representatives of a tradition of directing attention to problems of methodology in psychology, which goes from John Watson (1913; 1928), the founder of methodological behaviorism, to Paul Meehl (1967; 1991). Discussing the methodological problems arising as a result of the advent and use of novel research tools is a theme continually inherited by generations of psychologists. It

[1] Throughout this dissertation, I use the term "functional neuroimaging (abbreviated fNI)" to refer to all techniques of functional brain imaging, including positron emission tomography (PET) functional magnetic resonance imaging (fMRI), magnetoencephalography (MEG), etc. I use "PET" or "fMRI" when I talk about a specific imaging technique.

seems that this theme and the more general question of inference in psychological science deserve a fresh analysis by contemporary philosophers of science. The debate on the use of fMRI in psychological science provides a particularly fruitful example for a philosophical account of inference in psychology. This is because fMRI integrates essential methods of psychological science, namely behavioral experiments, use of statistical methods of data processing and analysis, and the employment of a complicated measurement tool to study relationships between mental processes and phenomena occurring in the human brain as subjects perform cognitive tasks. As witnessed by Machery (forthcoming), philosophers of psychology have ignored questions about inference and methods in their chosen science for too long. A careful philosophical study of fMRI methodology has the potential to shed light on the nature of inference and experimental knowledge in psychological science and will make a significant contribution to the philosophy of psychology. In addition, this kind of work will contribute to general philosophy of science as it discovers novel manifestations of well-known issues such as theory-ladenness, underdetermination, and Duhem's problem. Tackling these problems in the forms they arise in fMRI can give us fresh insights and help formulate novel construals of how we can deal with them.

In this chapter, I first provide a brief review of the literature on the epistemology of fNI. After covering the essentials of the literature, I point to some shortcomings of the works discussed. Then, I describe the approach I adopt, namely Deborah Mayo's error-statistical (ES) philosophy of science, in addressing problems that arise in fMRI research and argue that this account has the conceptual machinery that can resolve some of the inferential problems in fNI. After that, I discuss some telling examples of the

methodological problems in fMRI that are discussed by practitioners and statisticians who work on fMRI data. I also frame these problems in terms of the ES account of scientific inference and draw some conclusions about how the ES account can be employed to address these problems, which motivate and set the stage for the development of the project for this dissertation.

1.2: The Epistemology of Functional Neuroimaging

Since functional neuroimaging (fNI) is still a new field, there is not yet a sizable philosophical literature looking into the epistemological questions that arise in fNI. The philosophical work that has accumulated so far on this topic can be grouped into two categories: *Category One*: Discussions of the theoretical significance of fNI findings. *Category Two*: Discussions of the methodological difficulties of obtaining reliable inferences from fNI data. The work in category one predominantly focuses on the question of what we can learn from fNI results about human cognition and how, if at all, this knowledge can be used in evaluating theories of human cognition and adjudicating between competing theories. This kind of work generally fails to take notice of essential methodological aspects, the discussion of which is necessary for an assessment of the theoretical significance of fNI findings. The work in category two correctly emphasizes two major aspects of fNI research: 1. the high degree of complexity of the workings of fNI tools and experiments, 2. the immensely large data sets that fNI studies yield and the difficulty of statistically modeling these data and obtaining reliable inferences. Although the emphasis in this kind of philosophical work on fNI is right on target, unfortunately some of it falls victim to common misunderstandings and fallacious reasoning about

statistical methods. As a result, some of these authors end up giving in to skeptical arguments with the general conclusion that hypotheses of cognitive neuroscience are generally underdetermined by fNI data. If they are right, then scientists cannot learn anything from fNI results other than some broad heuristics for further studies. We start in the next section with a review and discussion of the philosophical work in category one.

1.2.1: Questions on the Theoretical Significance of fNI Findings: William Uttal (2001; 2002a; 2002b) has offered a strong criticism of functional neuroimaging research. Although he has raised important methodological issues characteristic of this field, Uttal's main criticism has been directed at the ways in which fNI results are taken to support modular theories of cognition. His central concern is whether or not we can accurately define isolated modules of cognition, which can then be localized in the brain using fNI techniques. One reason for this worry is the well-known difficulty of gaining access to cognitive processes occurring in the mind/brain for empirical study. Regardless of whether they are localized or parallel distributed in the brain, the inaccessibility of cognitive processes is perhaps *the* major methodological problem of inference not only in fNI but also in psychological science in general. A well-known discussion of this problem is an essay by MacCorquodale & Meehl (1948), which Uttal cites as support for his skepticism about fNI studies. He states that many cognitive processes that are assumed to separately exist by fNI researchers are in fact hypothetical constructs, which may not exist at all as they are described by the researcher. Because of this, Uttal concludes that the inferences a cognitive neuroscientist draws from fNI data reflects his or her prior theoretical commitments about which cognitive processes exist and how they

operate. Thus, regarding the cognitive processes that really occur in the brain, we may not be learning anything new from fNI results. Uttal also points out that neuroimaging tools cannot distinguish between different cognitive processes that occur in the same brain region. That is, the activation in a given brain region observed in a series of experiments may be due to different firing patterns of neurons in that region that realize different cognitive processes. Uttal suggests that fNI data can be misinterpreted as providing support for some modular theory of human cognition, while the possibility of different cognitive processes occurring in the same brain region may be overlooked. Because of problems like these, Uttal concludes that fNI data cannot be reliably used to adjudicate between competing theories of human cognition.

Bechtel (2002a; 2002b) agrees with the reality of methodological problems, but he rejects Uttal's conclusion. Bechtel reminds us that cognitive neuroscience is one of the disciplines that proposes and then tests theories that decompose a complex system into components. In this kind of research program, any and all such proposals are hypothetical; many of them are found to be incorrect and consequently revised. The important question, according to Bechtel, is not about the truth or falsity of modularist versus non-modularist theories, but whether or not cognitive scientists can obtain new findings from fNI that would be useful in evaluating their componential theories. New findings from fNI studies may support certain componential theories while refuting or necessitating revision in other theories. Also, Bechtel reminds us, cognitive scientists do not necessarily think that cognitive processes are directly accessible and they acknowledge that they have to rely on behavioral measures to study human cognition, for example, reaction time, performance on memory tests, etc. These behavioral measures

provide constraints for theories of cognitive neuroscience. Bechtel suggests that one reason why cognitive neuroscientists are eager to employ fNI tools is that data from these experiments can provide additional constraints on their theories which would help better evaluate existing theories of human cognition. Any theories that are incompatible with established fNI findings need to be discarded or revised. Therefore, the more useful approach to modularist theories is to see them not as final, complete theories of human cognition but rather as "fallible first proposals" that are subject to revision as we gather more data both from behavioral *and* fNI experiments.

In contrast to Bechtel, Hardcastle and Stewart (2002) agree with Uttal in denying the possibility of fNI being used to adjudicate between theories of cognitive neuroscience. They state that cognitive neuroscientists' modularist assumptions are "radically false" and add that they cannot get any support for their theories from fNI findings. This is because, prior to collecting data, researchers already assume the existence of specific cognitive processes, which are localized in different parts of the brain. They write; “It is a *bias* in neuroscience to localize and modularize brain functions” (p. S73). There are two central assumptions in this view; one is that different brain regions do different things. The other assumption is that the processes that are carried out take place in particular and confined streams. Hardcastle and Stewart make the strong statement that these assumptions are no more than mere prejudices which may “very well be wrong” (ibid.). They conclude that there is an inherent circularity in fNI research, which presumably destroys the potential of findings from this field to support modularist theories. Interestingly though, Hardcastle and Stewart also state that at this point in the history of psychology we simply cannot gather the kind of data that can

adjudicate between modularist versus non-modularist theories "for we have no way to approach studying the brain except through a modularist lens" (p. S80). This strong suggestion that the only conceptual frameworks for studying the brain are modularist in character does not leave many open doors for making progress on epistemological problems of fNI. (This is a topic I revisit at the end of chapter three.) Although Hardcastle and Stewart do not conclude that fNI research is useless, they suggest that fNI data have shown "precious little thus far" (ibid.). The rest of the debate on Uttal's criticisms revolves mostly around broad theoretical questions on modularity of cognition and localization of modules in fNI experiments and can be found in the second and third issues of the third volume of the short-lived journal *Brain and Mind*.

One expects complicated discussions of local issues in experimental methodology and inference from practicing cognitive neuroscientists because of their close proximity to the actual use of fNI techniques in their work. Practitioners should discuss specific methodological issues to a greater extent than philosophers of psychology, whose main interest has been in questions about large-scale theories of human cognition rather than establishing publishable findings. Several cognitive neuroscientists have written on inferential difficulties that arise in fNI, but their work also has focused on questions about the validity of inferences from fNI data to theories of cognitive neuroscience rather than inferences to real effects or specific experimental hypotheses. The debate between Richard Henson (2005) and Max Coltheart (2006; 2010) provides a good example; in essence, Henson believes that fNI results can be used to adjudicate between competing theories in cognitive neuroscience and Coltheart directly rejects this.

According to Henson (2005), fNI data comprise a dependent variable just like behavioral data we get in typical psychological experiments. The researcher adopts a systematic mapping between cognitive functions and brain structures. On the basis of the adopted mapping, Henson defines two types of inferences that can be drawn from fNI data; function-to-structure deduction and structure-to-function induction. The first type of inference we can draw when we do an fNI experiment with different experimental conditions (i.e. different cognitive tasks) and we obtain different patterns of brain activation in different groups. Based on this result, Henson thinks, we can infer that different cognitive functions are realized in different parts of the brain. The second type of inference can be drawn in the opposite direction: this is when we do an fNI experiment with different experimental conditions with different tasks but we obtain the same patterns of brain activation across groups. Based on this type of result, Henson suggests, we can infer that the same cognitive function is performed in different tasks.[2] Henson concludes that once we make the assumption that a systematic mapping exists between cognitive functions and brain structures and consequently we accept the validity of the two types of inference described above, then we can use fNI data to adjudicate between competing theories of human cognition just like we use behavioral data for the same purpose. This works when competing theories predict different patterns of brain activity in the same experiment and on the basis of fNI results we choose which theory has made the correct predictions. Henson (2005) discusses several examples of such tests of theories in fNI studies where researchers do the experiment and choose one theory over another. However, Coltheart (2006) discusses the same examples Henson uses as support

[2] This suggestion is subject to Uttal's criticism that fNI tools cannot distinguish between different patterns of activation in the same brain region that may give rise to different cognitive functions, see p. 8.

for his view and reaches a very different conclusion. Coltheart suggests that when we look at these examples in detail we realize that either the competing theories end up predicting the same pattern of brain activity, or the fNI study ends up yielding a simple result of localization of cognitive function, where the performance of a task is found to correlate with significant activation in a brain region. Coltheart does not question the validity of these kinds of inferences. For both Henson and Coltheart the major question is whether or not fNI results can be used to test competing theories. Coltheart looks at the literature so far and concludes that cognitive theories do not predict specific fNI results, because the terms of cognitive theories are underdetermined by behavioral and fNI data. Because of this, fNI experiments cannot be successfully used to adjudicate between competing theories. Henson and Coltheart are not the only researchers who disagree on the epistemic value of fNI results in the context of theories of cognitive neuroscience; Colin Klein (2010a) provides a review of similar works by practitioners in the field.

The themes in the above paragraphs remind one of good-old (or bad-old) familiar difficulties faced by the science of mind that have been extensively debated in psychology especially when it was just appearing as a separate discipline; e.g., the inaccessibility of mental phenomena. Thus, the underlying problems discussed above did not arise as a result of the advent of fNI techniques, but have been with us since the beginnings of psychological science. The point at which they cross paths with fNI is when we ask whether or not fNI data could be used to decide between competing theories of human cognition. Indeed, this is the major issue around which the discussion develops between Uttal, Bechtel, and Hardcastle and Stewart, as well as the one between cognitive scientists Henson and Coltheart. A shortcoming of this literature is that the debate has

been overly theory-centered in the sense that the major question has been taken to be whether or not fNI results are theoretically significant. If they are, then the question has become whether they can be used either as providing evidence (for or against) theories, or merely as additional constraints on theories of human cognition, which are desired to be neurally realistic.

However, there is a rich source of questions to be raised in the realm of fNI methodology and evidential import of fNI data for *experimental hypotheses* regardless of their significance for large-scale theories of cognition. In fact, all the above authors raise methodological issues as they discuss their theoretical concerns, but it appears that these issues are not taken to be philosophically interesting or fruitful. When we approach fNI from a primarily methodological and experimentalist standpoint, the important question becomes 'what exactly is the evidential import of fNI data regarding experimental hypotheses?' rather than 'what do fNI data show about theories of human cognition?' If we accept the claims of the skeptical authors discussed above, it appears to be the case that large-scale theories of human cognition are underdetermined by fNI data. But then, we can distance ourselves from large-scale theories and look at experimental knowledge in fNI and we can ask the question 'what is not necessarily underdetermined by fNI data?'Can we really not learn anything from fNI studies? What about experimental hypotheses tested in fNI studies independently of theory? If we cannot learn anything from fNI about experimental hypotheses either, then all the resources that go into fNI research is a terrible waste. Perhaps we can learn something from fNI, but what exactly? Such questions of underdetermination and evidential import of fNI data would take us to general questions about inference and experimental knowledge in cognitive neuroscience.

I continue in the next section with a discussion of the work that specifically discusses the methodology of fNI.

1.2.2: Discussions of the Methodological and Inferential Difficulties of fNI: Colin Klein (2010b) and Adina Roskies (2008, 2010) are two philosophers of science who have focused, in some detail, on the methodological aspects of fNI. Klein builds his skeptical argument regarding the epistemic value of fNI data by criticizing the use of statistical significance tests as they are employed in fMRI experiments. Roskies' argument, on the other hand, points to the great number of technical and inferential steps that have to be taken to prepare raw fNI data for statistical analyses, the results of which are then presented as neuroimages. These are the colorful pictures of the brain in action that are presented as fNI findings to scientists and the public alike. The complexity of these inferential steps and the fact that a great number of them are necessary, Roskies claims, lower the reliability of inferences drawn from fNI data. The idea behind this claim is that too often, we may obtain significant results not because there is a real effect but because we performed certain procedures on raw data which biased the analyses. Klein and Roskies have different arguments based on various premises, but they both come to similar skeptical conclusions. Although I wholeheartedly agree with the strong emphasis they put on questions of experimental methodology and inference, I disagree with their skepticism. Let us briefly review the central elements of their arguments.

Klein (2010b) recently raised criticisms that are directed at the use of statistical hypothesis testing, e.g., t-tests, in fMRI studies. Researchers generally compare the observed brain activation as measured by fMRI across conditions of interest. For

example, they compare brain activation observed in a control condition, where subjects do not do anything, with activation observed in an experimental condition, where subjects perform the given cognitive task. In terms of a simple t-test, they test the null hypothesis that there is no significant difference between control and experimental conditions against the alternative hypothesis that predicts a difference. The null hypothesis assigns probabilities to certain outcomes in the scenario where it is true. The probability of a certain outcome under the null hypothesis is its p-value. If the p-value of the observed outcome is smaller than a predetermined significance threshold, then we have a significant result; we reject the null hypothesis and conclude that there is a significant difference between the control and experimental conditions. The central premise in Klein's argument is the relative ease of finding significant results in fMRI studies. He gives some examples of how significant results can be obtained even when there is no real effect. For example, in order for a region of the brain to be identified as 'active' there has to be a statistically significant difference between observed activation in that brain region in different conditions of the experiment. Thus, choosing an overly liberal threshold for significance may lead to significant results. The charge is that when we observe significantly high activation in a given brain region, this may not be because there really is significantly increased activity in that region as a result of performing some cognitive task, but because we have chosen a significance threshold too liberal that it picks up background noise as if it is a real effect. Klein concludes that this arbitrariness negatively affects the reliability of inferences from fMRI data. This is a real problem and is known in the error-statistical literature as the simple fallacy of rejection. (It is also related to the problem known among fMRI practitioners as thresholding.) As discussed

by Mayo and Spanos (2006), one may have obtained a result that is beyond the significance threshold and thus deemed a statistically significant difference from the null hypothesis. However, one also has to consider the achieved p-value for the obtained result. That is, one has to consider how probable this specific result is, if there is no real effect. If this probability is high, then indeed, the statistically significant result may not support the inference that there is a real substantive effect, namely, significantly higher activation in the given brain region than would be expected if the null hypothesis is true. Reporting the achieved p-value instead of reporting the outcome in terms of a binary significant/not significant result may be a first quick way of addressing this problem. However, there are factors other than the chosen threshold that may bias analyses and yield significant results in the absence of a real effect. There are means to deal with these factors that are in line with Mayo's error-statistical (ES) philosophy of science; I address these and similar issues in chapter four where I discuss experimental and data models of fMRI experiments.

Another problem Klein (2010b) talks about is that the signal-to-noise ratio in fMRI can be improved by increasing the number of subjects in the experiment, which may increase the sensitivity of the experiment. Consequently, the experiment may yield significant results that may have occurred only because the number of subjects was increased. One central claim in Klein's account of fMRI is that these problems arise not as a consequence of the inherent characteristics of fMRI, but because of the nature of statistical hypothesis testing. He raises several other criticisms of significance tests and concludes that fMRI runs into problems because it requires statistical hypothesis testing to draw inferences about substantive hypotheses. I will argue that Klein's criticisms of

hypothesis testing stem from misunderstandings and misuses of these tests. As we will see in chapters three and four, the problems Klein discusses can be resolved when we approach them in terms of the error-statistical notions of severe tests and error probabilities. For example, the problem arising from increasing the number of subjects is called the large-N problem in philosophy of statistics ('N' is the abbreviation for sample size, i.e. number of subjects in an experiment). The large-N problem is not really a problem, because it can be solved by complementing statistical analyses with Mayo's notion of severity of tests, which differentiates between experiments with different numbers of subjects. As N gets larger, the variance of the data is reduced. Since the variance of the data is the denominator in the calculation of the test statistic, as N gets larger, mathematically the observed test statistic gets larger independently of the truth or falsity of the alternative hypothesis. Consequently, as N gets larger, it gets more probable to obtain a significant result in the absence of a real effect, that is, when the null hypothesis is true. Thus, in the error-statistical framework, a significant result is less indicative of a real effect if it was obtained in an experiment with a large N than in an experiment with a smaller N (Mayo, 1996; 2005b). In this way, we can distinguish between statistically significant fMRI results on the basis of the sample sizes of the experiments which yielded them, the larger the sample size the less indicative are the results of a real effect, because we may have obtained those significant results as a result of having a large sample size that makes for an oversensitive test picking up noise as if it is a real effect.

Adina Roskies is another philosopher of science who has emphasized methodological problems in fNI. She employs a distinction between the actual versus

perceived epistemic status of conclusions and suggests that the perceived epistemic status of neuroimages, i.e. the form in which fNI findings are presented, is higher than their real status (2008; 2010). In order to interpret results correctly, of course, we need to have a way of knowing the actual epistemic status. Roskies states that "determining actual epistemic status will involve a characterization of the inferential steps that mediate between observations and the phenomena they purport to provide information about. This characterization will include both the nature of the steps, and their relative certainty…" (Roskies, 2010; p.197). Roskies introduces the term *inferential distance* to refer to the totality of these inferential steps; the more the inferential steps the bigger the inferential distance. In the fMRI literature, some of these steps are referred to as preprocessing of data, but statistical modeling and analysis of data would also be included in what she calls inferential steps. As Roskies' diagnosis goes, the problem in the case of fNI is that there is a mismatch between the "actual inferential distance" and the "apparent inferential distance" between actual brain activity and the neuroimages that are presented as fNI findings. She writes, "I use 'actual inferential distance' to refer to the inferences explicitly employed in a scientific practice, while 'apparent inferential distance' indicates a more subjective measure characterizing the confidence people place in a conclusion on the basis of evidence" (ibid.). Roskies may be right in saying that people seem to overinterpret fMRI results. However, Roskies' account runs into problems when she makes the further assumption that inferential distance in fMRI cannot be characterized univocally. It is definitely a fact that there are a great number of technical and inferential procedures in fMRI experiments that have to be carried out between initial measurements of brain activation and final neuroimages. These steps require complicated computational

procedures on immensely large data sets. Because of the complexity of these procedures, Roskies concludes that the number and nature of these inferential steps cannot be sufficiently characterized (Roskies, 2010). She suggests that this lowers the reliability of inferences drawn from fMRI data, which leads her to a pessimistic conclusion about the epistemic value of fMRI findings. The crucial thing to note here is that Roskies' "inferential distance problem" can be satisfactorily addressed when we apply the hierarchical framework of models of inquiry to fMRI. I describe how this can be done in detail in chapter three, but briefly stated, we can break down an fMRI study into its component parts from experimental design to initial data collection, and from preprocessing of raw data to statistical modeling and analysis. We can then assess the error characteristics associated with each component. This enables us to assess on a case by case basis whether or not a given experiment constitutes a severe test of the specific inference that researchers intend to draw from the data. For example, let us say that an experiment yields a significantly higher activation in a certain brain region between control and experimental conditions. On the basis of this result, researchers conclude that the experimental treatment, say a cognitive task subjects performed, led to this result and make the substantive inference that the brain region in which higher activation was observed is involved in the performance of the cognitive task subjects performed. Now, we can carry out error-statistical analyses of this experiment looking carefully into the error characteristics or probabilities of its component parts. If these error probabilities are high, then we may not have evidence for the researchers' substantive conclusion. This is because, the significant result may have been obtained due to a bias introduced by a component part of the experiment. If the error probabilities are low enough to rule out or

minimize biases or errors, then we can safely conclude that the researchers have support for their substantive inference. In this way, we can test whether or not a given experiment constitutes a severe test of the specific substantive inference of interest. This is how we can "go the inferential distance," as it were.

In this section, I have provided a general review of the essential points of the philosophical literature on the epistemic value of fNI as an investigative tool of cognitive neuroscience. I begin the next section by pointing to some shortcomings of this body of work. Then, I describe the approach I adopt, which can remedy some of these shortcomings.

1.2.3: Some Shortcomings of the Philosophical Literature on fNI: I identified two major shortcomings of the body of works discussed above. One is that the majority of the work has been too theory-centered in its approach. We have seen this in the works of Uttal (2001), Hardcastle and Stewart (2002), and Bechtel (2002a, 2002b) as well as cognitive neuroscientists Henson (2005) and Coltheart (2006). This theory-centered approach is not very useful for two reasons. First, the problems about the theoretical significance of fNI data cannot be resolved without careful and detailed scrutiny of the methodological characteristics of fNI studies. A lot of the skepticism about the theoretical significance of fNI data is based on arguments that point to the unreliability of inferences in fNI studies. But this is a methodological problem and an account that specifically addresses aspects and problems of inferential procedures is required to address it. Skepticism about the epistemic value of fNI studies cannot be addressed without going into the characteristics of the methodology of fNI (e.g. PET or fMRI); only in those

details one can begin to address worries about unreliable inferences. The second reason that the theory-centered approach is not useful is that when we look at fNI studies only as potential sources of evidence for or against theories of cognition, we overlook philosophical questions that arise in fNI about experimental knowledge and inference. Approaching the epistemology of fNI from a theory-centered perspective precludes the possibility of tapping this rich source of fruitful questions, the answers to which may shed light on general questions regarding the kind of knowledge we gain from fNI studies and the nature of inference in cognitive neuroscience.

The other major shortcoming of the philosophical literature on fNI can be found in the smaller part of the literature that points to methodological problems of fNI studies. The works of Klein (2010b) and Roskies (2008; 2010) are in this group. They correctly identify some of the methodological difficulties, such as problems in the use of statistical significance tests (Klein), or the great number of preprocessing steps necessary before fMRI data can be analyzed, which are claimed to lower the reliability of inferences (Roskies). The major assumption in these works is that these methodological difficulties cannot be satisfactorily resolved. This assumption is false, because, as we will see below and in chapters three and four, error-statistical notions of error probabilities and severity of tests, together with the hierarchical framework of models of experimental inquiry, can be employed to resolve the problems raised by both Klein and Roskies as well as others (for example, Bogen, 2002). I adopt an approach that puts the emphasis on the experimental knowledge we gain from fNI studies rather than what fNI findings mean for large-scale theories of human cognition. It motivates specific questions; such as, what kinds of hypotheses can we reliably test in fNI experiments? How do we obtain reliable

inferences despite many sources of error? What kinds of inferences are warranted by fNI data? I apply Deborah Mayo's error-statistical (ES) philosophy of science (1996; 2005a) to address these questions.

1.3: The Error-Statistical Account and Severe Tests

In general, in order to address the problem of evidential import of data, we need to first know, for any scientific hypothesis, under what conditions we can conclude that we have evidence for it. We can start with a weak requirement: for an experiment or test, the weakest requirement is that it should not be guaranteed to find evidence for some effect regardless of whether or not there is a real effect. This gives us what Mayo and Spanos (2010) call the weak severity principle: Data **x** do not provide good evidence for a hypothesis *H* if **x** result from a test procedure with a very low probability or capacity of having uncovered the falsity of *H* (even if *H* is incorrect). This notion is the fundamental basis of the account that scrutinizes experiments by analyzing them with respect to their error probabilities – this is what Mayo calls the error-statistical account. Error probabilities provide the information on how frequently methods of research can discriminate between alternative hypotheses and how reliably they can detect errors. In light of these concepts, we can better address the question 'when do data **x** provide good evidence for a hypothesis *H*?' To do this, we can take Deborah Mayo's full severity principle as a guide, which states: "Data **x** (produced by process *G*) provide a good indication or evidence for hypothesis *H* (just) to the extent that test *T* severely passes *H* with **x**" (Mayo, 2005a; p.100). For a hypothesis *H* to pass a severe test *T* with **x**, two things must obtain; *first*, data **x** fits or agrees with *H*, and *second*, test *T* would have

produced, with high probability, data that fit less well with *H* than **x** does, were *H* false (Mayo, 1996; Mayo, 2005a). The idea here is that data **x** is evidence for hypothesis *H* just to the extent that the accordance between **x** and *H* would be difficult to achieve were *H* false. In other words, one must have done a good job at probing the ways one may be wrong in inferring from an accordance between data **x** and hypothesis *H* to an inference to *H* (as well tested or corroborated). It is important to note here that the severity of a test is not a feature of only the test itself. Severity assessments are carried out always on a specific test *T*, with specific test result $\mathbf{x}_0$ and a specific hypothesis *H*, so it is a function of three things, the *test*, (or the experiment); the *data*; and the specific *hypothesis* about which an inference is drawn (Mayo, 2005a). We can use the abbreviation SEV(*T*, $\mathbf{x}_0$, *H*) to mean "the severity with which *H* passes test *T* with $\mathbf{x}_0$" (Mayo & Spanos, 2006; 2010), where the severity function SEV(*T*, $\mathbf{x}_0$, *H*) can be calculated to get a quantitative value between 0 and 1. Although this is mainly about experiments and statistical tests, the notion of severity can also be employed in discussing error characteristics of experimental tools. In any given experiment, in order to assess whether or not the experiment constitutes a severe test of the hypothesis of interest, we need to know the error characteristics associated with the components of that experiment, such as instruments used for measurement and data collection, processing of data and statistical modeling and analyses.

In the context of fMRI, the question of evidence becomes 'when do we have evidence for a hypothesis that we want to test in an fMRI experiment?' In fMRI experiments, most of the time the hypothesis of interest predicts the effect of higher activation in some brain region in response to a cognitive stimulus or task. Let us call this

a real effect hypothesis. Armed with the notion of severe tests we can assess any given fMRI experiment with respect to whether or not it has put a specific real effect hypothesis to a severe error probe. If it has not, that is, if it was a test of low severity, then its results are evaluated accordingly. The results of an experiment of low severity cannot provide evidence for the hypothesis the experiment was meant to test. If the experiment yields a result fitting the real effect hypothesis, that is, higher activation in the brain region of interest, this does not constitute evidence for the hypothesis. Since the experiment was a test of low severity, it would have, with high probability, yielded a fitting result even if the real effect hypothesis is false. But in cases where an experiment of low severity does not yield results fitting the real effect hypothesis, this can be taken as evidence against this hypothesis. This is because; the experiment would have yielded results that fit the real effect hypothesis even if it was false. Yet, if it still does not yield such positive results, then this provides evidence against the real effect hypothesis. The concept of severity enables us to properly assess the epistemic value of all those cases of fMRI studies that suffer from problems discussed by Klein, such as arbitrary thresholds or too large samples. If, for instance, the sample size of an experiment is too large and makes it yield, with high probability, results that fit the real effect hypothesis of interest even in the absence of real effects, then the experiment constitutes a low severity test of the real effect hypothesis. As such, we would not make the mistake of taking its results as evidence for this hypothesis.

If we recall the quote above from Paul Meehl, his main worry was that as experimental methods and instruments of psychological science improve, it seems that it becomes easier to obtain experimental results that agree with the alternative hypotheses

in statistical hypothesis testing. Of course, Meehl did not have fNI tools in mind when he wrote about this problem in 1967 as no fNI techniques were available back then. However, he was right in thinking that such easy-to-obtain results did not provide strong evidence for substantive hypotheses and it is quite striking how relevant Meehl's worries are today. Meehl's worries are quite at home in ES terms. If it is really the case that as experimental methods of psychology improve, it becomes easier for statistical alternative hypotheses entailed by theories of psychology to be supported by data, then we can say that tests of psychological theories that entail alternative hypotheses in significance tests become less and less severe. In other words, psychological theories pass tests of low severity, so results of significance tests that agree with the statistical alternative hypothesis do not constitute evidence for these theories. This is a kind of fallacy of rejection I discussed in section 1.1 above.

Colin Klein (2010b) recently raised similar doubts, specifically about the evidential import of data from fNI experiments for substantive hypotheses. One major premise of Klein's argument is the same complaint as Meehl that almost all null hypotheses in fMRI experiments are false because of the high causal density of the human brain. That is, in the brain everything is connected to almost everything else and, as a result, almost any cognitive task would cause significant activations across different parts of the brain, areas which may not in reality be directly involved in the performance of that task. Thus, Klein concludes that neuroimages cannot be evidence for or against substantive functional hypotheses, because significant results would occur regardless of the truth or falsity of those hypotheses. In ES terms, the skepticism again seems to be that fMRI experiments cannot put substantive hypotheses to severe tests; hence fMRI data do

not carry much evidential import for such hypotheses, if they carry any at all. Although the brain *is* a causally dense system, the problem Klein points to does not stem only from this fact about the brain; part of the problem is that some fMRI scanners may be calibrated to be oversensitive to detect as real effects small activations due to noise compared to the baseline activity of the brain. This is something, an error-statistician would say, that makes for a low severity test of real effect hypotheses. So, it is another example of a problem that arises in fNI research which can be addressed by the ES account using the notion of severe tests and error probabilities (I discuss the details of how this can be done in chapters three and four).

The ES account can be employed to address problems that arise in fNI other than problems of oversensitive tests or thresholding that make for low severity tests of real effect hypotheses. For example, at first look, the high complexity of the fMRI experiment as a whole, with all the inferential steps between raw data and final analysis, make it look as if it is very difficult, if not impossible, to assess the error probabilities associated with parts of any given experiment. But we need not be intimidated by the "inferential distance." If, as stated above, the severity of a test is a function of three things; namely, the *experiment*; the *data* obtained in the experiment; and the specific *hypothesis* of interest, then, in order to assess severity, we need accurate characterizations of each of these three aspects of any given experiment. We have to know what hypothesis is *meant* to and *can* be tested, how the experiment is carried out to generate data, and what the data look like. To achieve this, we have to look at experiments in terms of the models that connect the primary scientific hypothesis or question being investigated to the detailed procedures of data generation and analysis.

We can define, for any experimental inquiry, three types of models; models of primary scientific hypotheses, models of experiment, and models of data. These models help us clearly describe the local procedures that are required to establish the connection between raw data and the substantive hypotheses of interest. We can then break down any given experiment, on a case by case basis, into its primary models, experimental models, and data models. This gives us the hierarchical framework of models of inquiry, as described in Mayo (1996) and this framework can help us see how, if at all, any procedures of the experiment influence certain experimental outcomes and in turn affect the probativeness of the experiment. We can also find out if there are any experimental sources of error that need to be taken into account. Once we break an fMRI study down to its component parts, placing each one in its proper place in the framework of models, we first clarify the specific hypothesis that is tested by the experiment. Then, we can look at error probabilities that are associated with the component parts of the experiment that go into the experimental and data models. To what extent do they affect the probabilities of obtaining certain experimental outcomes? In some cases, we may have to see whether or not we even know the error probabilities to control for certain types of errors; if not, then we have to find out those error probabilities. The application of the hierarchical framework of models to fMRI can help us assess the reliability of inferences drawn from fMRI data. It can give us the kind of characterization of inferential distance, or something very close to it, and the "inferential distance" can be trodden safely, which, if we recall from above, is something that Roskies thinks can never be done. Therefore, by applying the ES account to fMRI, we can address not only the question of the evidential import of fMRI data by identifying the primary hypothesis of the given experiment, but we can also

identify the conditions that need to be met to conclude that we have evidence for that hypothesis.

So far in this section, I have described the central ES notions of severe tests, error probabilities, and the hierarchical framework of models of inquiry and I have discussed how they can be employed in addressing, hopefully resolving, the issues raised by philosophers who are skeptical about the epistemic value of fNI studies. In the next section, I will discuss some problem cases that have come about in the practice of fNI discussed by practitioners in the field. These cases provide concrete examples that will help me illustrate how the ES approach can address the methodological problems that are faced by fNI researchers.

1.4: Salmon Thoughts and Voodoo Correlations

Let us start with a concrete example of a typical fMRI study; cognitive neuroscientists Canli and colleagues (2002) studied the amygdala as a critical brain structure for the processing of emotional stimuli. In this study, subjects were given a personality test and then assessed for the personality trait, extraversion. After this, they were shown pictures of emotional faces (emotions displayed were angry, fearful, happy, and sad) or neutral faces as their brain activity was measured by the fMRI scanner. After data collection and preprocessing, the researchers analyzed the data to see if there were any significant differences in the subjects' brain activity between experimental conditions, i.e., when emotional faces were shown versus when neutral faces were shown. The results showed that all subjects had significant activity in the amygdala when they were shown fearful faces. Also, there was a significant positive correlation between subjects' scores on the

extraversion part of the personality test and activation in the amygdala in response to happy faces, so extraverted people had a higher degree of activity in their amygdala when they were shown happy faces in comparison to less extraverted people. The authors concluded that this brain mechanism, observed in fMRI as high activation in the amygdala when people were shown happy faces, may "contribute to behavior consistent with the sociable interactive style of extraverts" (p. 2191). This is an interesting finding, as are many other findings from fMRI studies, but the question I want to ask is how can we assess whether or not this is a real effect and not an artifact of the experiment? Of course, the same question can and should be asked about any other fMRI findings. An interesting demonstration by another practitioner of fMRI further motivates this question.

1.4.1: Salmon Thoughts: The neuroscientist Craig Bennett and his colleagues have carried out a telling demonstration about the possible misuses of the fMRI technique as a tool of empirical research (2010). They put a mature Atlantic salmon in an fMRI scanner, who was then shown photographs of individuals whose faces showed different emotions. The salmon was then "asked" to determine the emotions in the photographs. The results showed that several parts of the salmon's brain cavity were detected to be active when emotional photographs were shown compared to when nothing was shown. During scanning the salmon was not alive. So, what can we conclude from these results? Do these results show that a deceased salmon is cognitively capable to process pictures of emotional faces? If not, then what do they show, if anything at all, about the methods and analyses in fMRI studies? Without any doubt, the one thing that the salmon experiment demonstrates is that it is pretty easy to get significant results out of random noise in fMRI

experiments. For certainly the dead salmon did not process pictures of faces to identify the emotion shown. Does this mean that skeptics like Uttal, Hardcastle and Stewart, Klein, and Roskies are right in questioning the epistemic value of fMRI results? No, at least not necessarily so, because when we scrutinize this experiment from the perspective of severe tests and error probabilities, it can be straightforwardly seen that this experiment does not show anything like a dead salmon's cognitive capacity to process emotional faces. This is because the salmon experiment had little chance of finding evidence against the real effect hypothesis that the dead fish cognitively processed the pictures. So, the probability of erroneously rejecting the null hypothesis was high. In other words, this experiment had very high error probabilities; with high probability the experiment would have yielded positive results even if the null hypothesis is true. Thus, it was a low severity test of the hypothesis of interest, and accordingly its results would not be taken to support a real effect hypothesis by the error-statistician. But, at the metascientific level, the salmon experiment does point strongly to the need for careful scrutiny of fMRI as an experimental methodology from the ES perspective. Of course, talking about the salmon experiment is easy; it is pretty obvious that the deceased salmon did not engage in any cognitive activity. But what do we do when we get experimental data in support of real effect hypotheses that may be true? How can we establish that a given fMRI data set shows that there is a genuine effect and that the results are not due to an artifact out of random noise? For example, how can we establish as a genuine finding the results of Canli et al.'s study described above, which showed that extraverted people had higher activity in their amygdalae when they were shown pictures of happy faces?

Again, the notions of severe tests and error probabilities would come to our rescue. Let us now look at another case, which is not as obvious as the salmon experiment.

1.4.2: Voodoo Correlations: In most experiments we do not have complete data, and we face uncertainties between what we want to measure and what we can measure. We can study only limited samples of populations. In these contexts, we use statistical models to get evidence and draw inferences from data. One issue that arises is whether statistical results are due to mere chance. Statistical reasoning is supposed to help us in finding out what mere chance effects look like. In 2009, an article in Newsweek reported a debate about the use of statistical analyses in just the kind of fMRI studies that seem to provide supporting evidence for various hypotheses that may be true (Begley, 2009). This debate has caused considerable controversy in the fMRI world and beyond. The controversy started when cognitive neuroscientists Edward Vul and his colleagues looked at some correlation coefficients, reported in peer-reviewed scientific journals by fMRI researchers working in the field of social neuroscience, which seemed to be too high to be true. They raised this concern in a manuscript entitled "Voodoo Correlations in Social Neuroscience," which was accepted for publication by the journal *Perspectives on Psychological Science*. The manuscript became available online in several websites and was reported and commented on in the news media. After much attention and complaints by some fMRI researchers, before it was published, the title of the article had to be changed to "Puzzlingly High Correlations in fMRI Studies of Emotion, Personality, and Social Cognition."

Vul and his colleagues (2009) voiced their doubts about the overly high correlation coefficients by first defining a puzzle. They state that in psychometrics there is a statistical fact that the strength of the correlation observed between two measures A and B reflects not only the strength of the relationship between the traits underlying A and B, but also the reliability of the measures A and B (e.g., as would be assessed by applying test-retest reliability measures). In general, this relationship holds:

$$R_{ObservedA,ObservedB} = r_{A,B} * \text{sqrt}\,(\text{reliability}_A * \text{reliability}_B)$$

where $R_{ObservedA,ObservedB}$ is the observed correlation between scores on measures A and B, $r_{A,B}$ is the strength of the relationship between the traits underlying A and B and reliability_A and reliability_B are the reliability of the measures A and B, respectively (Nunnally, 1970; cited in Vul et al., 2009). On the basis of this fact, Vul et al. state that "the reliabilities of two measures provide an upper bound on the possible [meaningful] correlation that can be observed between the two measures" (Vul et al., 2009; p. 275). Vul and his colleagues suggest applying this rule to fMRI studies, which have been reported to yield very high correlations between scores on personality measures and certain patterns of brain activity as measured by fMRI. To do this, one needs the reliabilities of the fMRI procedure as a tool of data collection as well as the personality measures used in these studies. They cite several studies on the reliability of some common measures of personality such as the Big Five personality factors and the Minnesota Multiphasic Personality Inventory. On the basis of these reliability studies they reach the "optimistic estimate" of a range of .7 - .8 for the reliability of measures used in the fMRI studies that reported the high correlations. Vul et al. also report the results of some recent studies on the reliability of fMRI (Kong et al., 2006; Manoach et

al., 2001; Aron et al., 2006; Jonhstone et al., 2005; all cited in Vul et al., 2009). On the basis of the results of these studies they estimate the reliability of fMRI measures to be at or below .7 (ibid.).

Armed with reliability estimates for fMRI and personality measures, Vul and his colleagues define their puzzle. In the best case scenario, measures of personality have a reliability of .8 and fMRI has a reliability of .7, so we can plug these values in the equation above. That is, if we assume that an fMRI study is done, with no measurement errors, on some personality trait the measure of which correlates perfectly with activation in the associated brain region (i.e. r =1.0), then the highest possible meaningful correlation between scores on the personality measure and brain activity as measured by fMRI that could be obtained would be sqrt (.8 * .7), or .74. Recall that this is the value we would get in the ideal case, where the reliabilities of personality measures and the fMRI procedure are at the highest possible value and the underlying correlation between the measured trait and the pattern of brain activity associated with the trait is perfect. Of course, in actual studies, this state of affairs is not probable. So what is the puzzle? It is the fact that correlations exceeding this upper limit of .74 are reported in fMRI studies on emotion, personality, and social cognition published in peer-reviewed journals. If the upper limit for any meaningful correlations is .74, then what do these correlation coefficients much greater than .74 mean?

These fMRI studies that Vul and his colleagues (2009) have focused on were described in original articles that reported correlations between participants' scores on a behavioral measure and a statistic reflecting the brain activation within a given set of voxels (i.e. three dimensional volume elements corresponding to parts of brain tissue).

Vul and colleagues were interested in how these overly high correlations were calculated. When they studied the methods sections of these articles Vul and colleagues found that the statistical analyses greatly varied from study to study and also a lot of the analyses were not made clear, so they sent surveys to the authors of these articles regarding the methods used. The critical question was how were the voxels selected the fMRI data from which were correlated with scores on personality measures? Were they selected according to anatomical or functional criteria, or both? Were they selected because they were in or near a prespecified brain region (anatomical criteria), or were they selected, because they were found to be active in fMRI scans as subjects performed some task or did a behavioral test (functional criteria)? Or was the selection criterion a combination of anatomical and functional constraints. Another crucial question was if functional data were used to select the voxels, were they the same functional data as were used to define the reported correlations?

The results of the survey showed that 53% of the respondents said that "regression across subjects" was used as a functional constraint to select voxels "indicating that voxels were selected because they correlated highly with the behavioral measure of interest" (Vul et al, 2009; p. 278). Furthermore, it was found in the survey that all these studies used "the same data to compute the correlation as were initially used to select the subset of voxels" (ibid.). Thus, the procedure in these studies went like this: researchers ask each subject to do a behavioral test, so they obtain a behavioral measure from each subject and they obtain fMRI measures from many voxels as subjects do the test. Then, the fMRI activity observed in each voxel is correlated with the behavioral measures from all subjects; this gives them thousands of correlation coefficients. After

this, those voxels, for which the correlation passes a statistical significance threshold, are selected. In the final stage, the fMRI signal is aggregated across the selected voxels and a final correlation coefficient is calculated between the fMRI signal and subjects' scores on the behavioral test. Of this procedure Vul and his colleagues state: "Such an analysis will inflate observed across-subject correlations, and can even produce significant measures out of pure noise. … With enough voxels, such a biased analysis is guaranteed to produce high correlations even if none are truly present" (Vul et al, 2009; p. 279). They refer to the methodological fault in this procedure as the "non-independence error," which Vul and colleagues describe as selecting one or more voxels based on a functional analysis, and then reporting the results of the same analysis and functional data from just the selected voxels. They write: "This kind of analysis distorts the results by selecting noise exhibiting the effect being searched for, and any measures obtained from such a nonindependent analysis are biased and untrustworthy" (ibid.). Vul and his colleagues conclude that because a sizable group of researchers use the above procedures, a "segment of fMRI research on emotion, personality, and social cognition is using seriously defective research methods and producing a profusion of numbers that should not be believed" (ibid.).

The above conclusions are serious in the sense that they call into question the very dependability of a scientific field that is still in its developing stages. If the conclusions of Vul and colleagues are correct, they would have serious consequences for the foundations and future of cognitive neuroscience. Partially because of this and partially because of the controversy this manuscript caused with its original provocative title including the term "voodoo correlations," the editors of the journal *Perspectives on Psychological Science*

invited practitioners and statisticians of fMRI to contribute commentaries on Vul et al.'s paper. Several authors agreed and the contributed commentaries were published together with the original paper as well as replies by Vul and colleagues in the same issue of the journal. The end result was a debate, which illustrates the state of affairs in methodological discussions of fMRI. Here, I will first briefly review some of these commentaries that together represent the character of the overall debate and then I will discuss what the error-statistical approach to this debate would be.

The commentators' arguments and conclusions about the Vul et al. paper vary considerably. Nichols and Poline (2009) state that Vul et al.'s arguments come down to two points: 1. there indeed is a massive multiple-testing problem in fMRI, which is related to the validity of corrected versus uncorrected inferences; 2. the methods sections in fMRI articles are incomplete and/or confusing. Because of this, they conclude that Vul and his colleagues' thesis is overstated and the problems they point to are straightforward and can fairly easily be solved. Yarkoni (2009) thinks that the primary conclusion of Vul and his colleagues is correct, but disagrees about the reasons that lead to it. Yarkoni suggests that correlations in cognitive neuroscience are inflated and perhaps to a greater extent than Vul et al. described. But, Yarkoni states that this is not due to the nonindependence error as they define it; the primary cause of the inflated correlations is lack of statistical power in fMRI studies.

In contrast to Nichols and Poline and Yarkoni's commentaries, Lieberman, Berkman, and Wager (2009) directly challenge Vul et al.'s conclusions. They claim several of Vul et al.'s premises are incorrect: first, according Lieberman et al., whole-brain regressions, which Vul et al. say lead to spurious correlations, are a valid method of

identifying brain regions that have reliable correlations with individual difference measures. Second, typical fMRI sample sizes will only rarely produce large correlations in the absence of any true effect. Third, the magnitude of the bias is more modest than Vul et al.'s estimates. Because of these reasons, Lieberman et al. conclude that "correlations in social neuroscience aren't voodoo" and they add that Vul and his colleague's paper is a direct attack on the field of social neuroscience with an overly aggressive tone.

Nicole Lazar (2009), a statistician specializing on modeling and analyses of fMRI data, agrees with Vul and colleagues that the problem is real, but she identifies it differently from them. According to Lazar, the problem is selection bias, so it is neither a new problem nor a problem unique to fMRI. She writes that the "use of massive, complex data sets (common in modern applications) to answer increasingly intricate scientific questions presents many potential pitfalls to valid statistical analysis" and suggests that there is no immediate and straightforward solution (ibid., p. 308). Lazar suggests that new statistical methods are needed for the analysis of large data sets and calls for a "strong collaboration between statisticians and scientists and the development of statistical methods specific to the types of data encountered in practice" which can help researchers solve these problems (ibid.). In another commentary, Lindquist and Gelman (2009) agree with Lazar that the problem is real, but they claim that it cannot be resolved by standard statistical means.

As can be seen above, the commentaries on Vul and his colleagues' methodological puzzle vary considerably; some commentators think that there is no problem at all, some others think that there is a problem but not a serious one, while some

others think that the problem is real but cannot be resolved by any existing statistical means. In addition, those who believe there is a real problem disagree in their suggestions on how to proceed in addressing it. Given these commentaries, it is clear that not all of the commentators can be right. It appears that in addition to disagreement, there also seems to be confusion about basic terms and arguments. Those, who think there is a real problem, formulate it differently; these various formulations reflect different construals of the general issue. For example, Vul and his colleagues think the problem is that of nonindependence in the calculation of correlation coefficients, while Lazar thinks it is a standard case of selection bias and Yarkoni thinks the problem is not having enough statistical power. In this environment of confusion, the error-statistical account can help.

1.4.3: What Would the Error-Statistician Say? Research hypotheses to be tested in an experiment (broadly defined) are formulated in terms of a statistical model, which approximately, or ideally, represent the underlying data generating mechanism; in the case at hand, the whole fMRI experiment. If we want to test a hypothesis with an experiment one must have done a fairly good job at probing the ways one may be wrong in inferring from an accordance between data **x** and hypothesis *H* to an inference to *H* (as well tested or corroborated). In any given experimental context, we need to ask the question 'do any of the procedures influence the error probabilities associated with this experiment to make it highly probable to yield data that fit the hypothesis being inferred even if it is false?' The same question can be asked in a different way; 'does our experiment do a good job of testing for the ways in which the hypothesis may be false?' It is precisely this kind of question that we must ask about the experiments Vul and his

colleagues describe that report overly high correlations and create the methodological puzzle. What is the probability that the procedure applied in these experiments would yield a high correlation between fMRI data and the given behavioral measure, if in reality this correlation is very low or is equal to zero? According to Vul et al.'s analysis, the answer to this question is "very high." This procedure does increase the probability of erroneous inferences in any fMRI experiment that employs it. Therefore, any fMRI experiment that employs this procedure makes for a low severity test of the real effect hypothesis which in this case predicts a high correlation between activation in certain brain regions and scores on psychological tests. The reason for this is that, with high probability, the experiment would yield data that agree with the hypothesis even if it is false. Therefore, Vul and his colleagues are right in saying that the correlations coming from fMRI experiments that used this procedure do not constitute good evidence for the hypotheses that they were designed to test.

This was also the problem in the salmon experiment described above. If an fMRI experiment does no better than the salmon experiment in controlling for errors, then the data it yields is certainly no evidence for any hypothesis. Error-statistical scrutiny enables us to find out whether an experiment does a good job of controlling for errors and the debate on voodoo correlations is precisely about this question. Part of the problem stems from the fact that our research tools in contemporary science are highly complicated. As we employ these tools, philosophical problems arise, such as the notion of a valid scientific inference in the face of uncertainty, variability, and error. In the debate on voodoo correlations, the commentators propose different solutions and prohibitions, and one reason for this may be that they do not have a solid philosophical notion of valid

scientific inference, which can be used to put everything in a coherent framework. In general, it appears that the real issue is about what effect a given experimental procedure has on the error probabilities of the fMRI experiment. Do they make it highly probable to find a significant result even if there is no real effect? It is this question that must be asked about every procedure of the fMRI experiment, from experimental design to collection and preprocessing of data to statistical modeling and analysis.

Researchers in fMRI readily acknowledge that fMRI experiments involve immensely huge data sets (measurements of fMRI signal coming from between 40,000 and 500,000 voxels), several 'preprocessing' steps to reduce noise and prepare the data for comparisons across different brains, and then procedures for obtaining contrast measures of fMRI signals across different tasks, modeling data and running statistical analyses on the data to infer whether or not there is a significant difference between different tasks. What effects, if any, do these procedures have on the error probabilities of any given fMRI experiment? The editor of *Perspectives on Psychological Science*, the journal that published Vul et al.'s original paper and the commentaries on it, suggests that fMRI is a field that is faced with "somewhat unique" methodological problems (Diener, 2009). This is certainly true. Yet, as Lazar (2009) notes, there are some problems that are not unique to fMRI, for example, multiple testing problem. However, Vul and his colleagues and some of the commentators rightly state that standard statistical solutions to such problems do not work. At the same time, there are some problems that *are* unique to fMRI and other techniques of neuroimaging. For example, the anatomical variability of brains across individuals, or how characteristics of the fMRI scanner used in an experiment influence error probabilities; e.g., scanners of higher magnetic field strength

have higher detection sensitivity and this may cause noise artifacts. Unique or not, we should inquire about the effect each of these problems has on the error probabilities associated with fMRI experiments. The error-statistical account has the conceptual machinery that can help us carry out such inquiries: we can apply the hierarchical framework of models of inquiry to fMRI studies, as discussed above, where different component parts of the study are placed in primary models, experimental models, and data models. Then, we can inquire into the error characteristics or error probabilities associated with these components and find out to what extent, if any at all, they affect the probability of obtaining certain experimental outcomes. For example, the procedures of voxel selection and statistical analysis described in the voodoo correlations case would be placed in the data models and scrutinized with respect to their error probabilities. (I revisit voodoo correlations in chapter four.) This kind of error-statistical scrutiny will help us identify the conditions under which we can carry out fMRI experiments that can reliably test the hypotheses of interest, carry out severity assessments of these experiments, and draw reliable inferences from fMRI data.

In addition to addressing problems in the epistemology of functional neuroimaging, the kind of philosophical work I do in this dissertation has the potential of contributing to general philosophy of science; it can help us learn about the nature of experimental knowledge in scientific fields that use highly complicated tools and techniques of data collection and analysis as well as the methodological problems that arise in these fields. Also, the work here can give us crucial insights about certain debates within philosophy of science when we rethink them in the context of experimental knowledge in cognitive neuroscience, for example Duhem's problem,

underdetermination, etc. Once we have a clearer and more complete account of inference and experimental knowledge in cognitive neuroscience, we can better evaluate the claims of theoretical accounts that use results from fMRI studies as evidence for or against various theories of human cognition, philosophy of psychology, and even ethics.

Before moving on, let me briefly give an outline of how I proceed in the rest of this dissertation. Knowing the history of fMRI helps with any analysis of fMRI. It is particularly helpful when one wants to carry out a conceptual analysis of the methodology of fMRI. Knowing how it was invented and on the basis of what scientific knowledge, namely physics of magnetic resonance and physiology of metabolism in the brain, helps us better understand the nitty-gritty details of fMRI as an empirical tool. This is central to doing fruitful methodological and inferential analyses. In addition, looking at the history of fMRI comes with a philosophical bonus; it provides an example in the context of which we can talk about theory-ladenness in a new light. A lot of the philosophical criticisms of fMRI are based on the notion that fMRI findings are theory-laden, but the term 'theory' in these criticisms refers to the modularist theories of psychology. The other kind of theory-ladenness that arises in fMRI is a useful kind of theory-ladenness in the sense that fMRI findings are laden with the theories of physics and physiology; that is they are arrived at by the knowledge that these other sciences have produced. Therefore, this kind of theory-ladenness allows for the representation of and intervention in psychological constructs using well-established concepts and methods of physics and physiology. This is similar to how Ian Hacking (1983) describes that the reality of dense bodies in blood is established using different types of experimental techniques, namely, electron microscopes and fluorescent staining. This can be seen as an

argument from coincidence; it would be a highly improbable coincidence that independent procedures yield the same result, small dots in red blood platelets, unless these dots are real entities rather than instrumental artifacts. In a similar way, representing and intervening in psychological constructs using fMRI, which depends on physical phenomena independent of psychology, provides additional support for the reality of these psychological constructs.

In chapter three, I begin with the essential question of what it is that we can learn from fMRI experiments despite the complexity of the methods and many sources of error. In order to answer this question, I use the notion of severe tests and the hierarchical framework of models of inquiry. First, I describe the hierarchical framework of models in detail as formulated by Mayo (1996). Then, I emphasize one part of the hierarchy, namely the primary models and discuss what can legitimately be placed in the primary models in fMRI experiments. This clarifies the evidential import of fMRI data; specifically speaking, we can learn about hypotheses about hemodynamic activity in the brain that is related to cognitive processes. Although this chapter gives us what we can, in principle, learn from fMRI experiments, it does not tell us how we can learn it, that is, how we can draw reliable inferences from fMRI data. In chapter four, I emphasize the other parts of the hierarchical framework of models; namely experimental and data models. Chapter four tells us how we can learn, that which we can learn from fMRI, by carefully scrutinizing the methodological and inferential procedures in fMRI studies from experimental design to data collection and modeling and analysis. I put everything together in the final chapter. Chapters two, three, and four offer an error-statistical epistemology of functional neuroimaging. In the final chapter, I discuss the implications

of this work for inference and experimental knowledge in cognitive neuroscience answering questions such as what kind of knowledge can cognitive neuroscience give us? How can it do so? What is the nature of the relationship between experimental knowledge in psychology and large-scale theories of psychology? Of course, I cannot definitively answer all these questions in my dissertation, but I argue that the work I do here can shed some light on these questions and lead us to novel insights about knowledge in psychological science.

Chapter Two:

A Concise History of fMRI: Theory-Ladenness of a Useful Kind

> We must suppose a very delicate adjustment whereby the circulation [of blood] follows the needs of cerebral activity. Blood very likely may rush to each region of the cortex according as it is most active, but of this we know nothing. I need hardly say that the activity of the nervous matter is the primary phenomenon, and the afflux of blood its secondary consequence.
>
> William James, 1890 (p.99)

2.1: Introduction

As can be seen in the above quote from William James' classic *The Principles of Psychology*, one of the founding figures of psychology, the idea that there must be a relationship between mental activity and patterns blood flow in the brain is not new. However, a century or so had to pass before we had a research technique that indeed enables us to observe and measure the strength of the relationships between cognitive activity and cerebral blood flow. In the century that passed, several scientific developments in different sciences took place, which finally gave us functional magnetic resonance imaging (fMRI), the essential research technique of the new field of cognitive neuroscience. In this chapter, I provide a historical account of these developments. Studying the history of the development of fMRI serves two purposes: first, it enables one to better understand the scientific and technical principles on which fMRI scanners are built and operate. This understanding will be of crucial importance in the rest of this dissertation, particularly in chapters three and four. Second, the historical account of the

development of fMRI provides an example of a scientific research paradigm in which the empirical needs of one discipline, cognitive neuroscience, are served by scientific knowledge from two other disciplines, physics and physiology. This makes it possible to discuss a kind of theory-ladenness of observations in cognitive neuroscience that is different from the kind often discussed by philosophers of psychology, as well as those who have written on the epistemic value of fMRI findings.

2.2: The Beginnings of Cognitive Neuroscience

As can be seen in the above quotation, one can find postulations about possible connections between mental activity and cerebral blood flow in sources as early and foundational as William James' textbook *Principles of Psychology*. As Raichle (1998) notes, James cites as support for this postulate the work of Italian physiologist Angelo Mosso (1881) who recorded brain pulsations of the cortical structure in three patients who had lesions of the skull, which enabled Mosso "to let the brain-pulse record itself directly by a tracing" (James, 1890; p. 98). Mosso found that blood flow in the brain rose when the subjects were spoken to or when they engaged in cognitive activity such as doing arithmetic mentally (ibid.). In the same year that *Principles of Psychology* was published, Mosso's findings and James' postulate were supported further by two British physiologists, C. S. Roy of Cambridge and C. S. Sherrington of Oxford, who proposed and provided experimental evidence for what neuroscientists today call the Roy-Sherrington hypothesis (Roy & Sherrington, 1890). This hypothesis stated explicitly that the volume of blood flow in the brain varies locally in parallel with changes in local functional activity. In other words, when a certain brain region becomes more active in

response to some stimulus or as it performs some cognitive task, such as walking or looking at a picture, the nervous system has mechanisms that increase blood flow to that region. Thus, variations in cerebral blood flow reflect variations in functional activity in the brain. This ostensibly simple hypothesis has formed the basis for the field of research that we today call cognitive neuroscience, which employs several different techniques of functional neuroimaging (fNI) with functional magnetic resonance imaging (fMRI) as its preferred tool of data collection. Yet, no less than a century of experimental research in diverse fields of science had to be done before the ideas and works of Mosso, James, and Roy and Sherrington could be applied in the new field of cognitive neuroscience. One reason for this was the lack of tools sufficiently developed for the study of relationships between mental activity and cerebral blood flow at the time. Another reason was the work of eminent British physiologist Leonard Hill (1896). Due to the lack of sufficiently developed tools, the experiments of Hill were inadequate to investigate relationships between cognitive activity and cerebral blood flow and Hill concluded that no relationships existed between cognitive activity and cerebral blood flow (Raichle, 1998). Because of Hill's prominent reputation as a physiologist, this conclusion thwarted further scientific development on this topic for decades. However, developments in physics and physiology scattered throughout the twentieth century eventually made it possible for scientists to study these relationships. Before moving on with the details of these developments, let us look briefly into how cognitive neuroscience came to be in the late twentieth century.

Raichle defines cognitive neuroscience as follows: "Cognitive neuroscience combines the experimental strategies of cognitive psychology with various techniques to

actually examine how brain function supports mental activities" (1998; p. 765). 'Cognitive Neuroscience' was a term that was coined in the late 1970's in discussions between Michael Gazzaniga and the prominent cognitive psychologist George Miller. Gazzaniga describes what they had in mind as follows;

> What he [George Miller] meant by cognitive neuroscience was to emerge – slowly. What we already knew was that neuropsychology was not what we had in mind. Tying specific functions to lesioned brain areas was not going to be our enterprise. The bankruptcy and intellectual impoverishment of that idea seemed self-evident, especially with the advent of new brain-imaging techniques that revealed how much else was always damaged following what had previously been thought to be focal damage (Gazzaniga, 2000; p. 4).

There are two points in Gazzaniga's above remarks worth mentioning. First, it is clear that Gazzaniga and Miller hold that the advent of new brain imaging techniques has shown the inadequacies of the old fashioned way of studying the relations between the mind and the brain, namely cognitive neuropsychology. These new techniques, such as fMRI, provided images of the brain with much higher spatial resolution and therefore enabled researchers to make much more accurate assessments of the damage in patients with brain injuries who have cognitive impairments. What was thought to be a good technique, that is, correlating the locus of damage in the brain and cognitive impairment, was shown to be unsatisfactory. If we do not have damage that is in a neuroanatomically well-defined structure or region of the brain, which happens very rarely with naturally occurring injuries, then it is problematic to correlate a certain cognitive function with a coarsely defined region in the brain.

The second thing to emphasize in Gazzaniga's remarks is how the capability of fMRI for noninvasively providing images of the brain hinted at the possibility of using these techniques not just for purposes of medical assessment but perhaps also for studying what happens in the brain as individuals engage in cognitive activities. If this is possible, then perhaps we could have a way of investigating the brain as it performs cognitive tasks with good spatial resolution. Furthermore, we would not have to worry about how precisely located is the damage in a patient's brain. Much better yet, with the new brain imaging techniques, we would not even have to wait for patients with brain damage; we would be able to study intact brains. The arrival of techniques like PET and fMRI provided all these advantages. In the next section, I provide a concise history of the development of fMRI, which became the major research tool of cognitive neuroscience.

2.3: History of fMRI

The history of the development of fMRI includes two separate trajectories of research and development; one in physics, which led to the discovery of the phenomenon of nuclear magnetic resonance (NMR) and its later application in the development of magnetic resonance imaging (MRI), the other in physiology, which involved the discovery that changes in blood flow in the brain could be measured using MRI and the description of the blood-oxygen-level-dependent (BOLD) response. We start with the trajectory in physics.

2.3.1: The Discovery of Nuclear Magnetic Resonance: In the 1920's, the Austrian physicist Wolfgang Pauli, having noted some anomalies in the electromagnetic spectra emitted by excited atoms, postulated that atomic nuclei had two properties, namely, spin

and magnetic moment (Huettel et al., 2008; p.15). These properties could take only discrete values; that is, atomic nuclei can spin only at certain frequencies and produce only particular magnetic forces. Because in the 1920's, the science and techniques to test these conjectures were not available, Pauli's ideas could not be put to the test until the 1930's. However, there was an early technique for the study of spin properties of atomic nuclei developed by the German physicists Otto Stern and Walther Gerlach. In this technique, a gaseous beam of a single element is sent through a static magnetic field and then hits a detector plate (ibid.; p. 16). There are two possibilities for what would be observed on the detector plate: 1) If atomic nuclei can spin only at a number of certain frequencies, then the magnetic field would split the beam into some finite number of smaller beamlets and several separate beams would be seen on the detector screen; 2) If, however, the spin frequencies of atomic nuclei can take a continuous range of values, then a continuous distribution of intensity would be observed on the detector screen. Quantum theory predicted that the beam would be split into discrete beamlets and this was exactly what was observed in the Stern-Gerlach experiment (ibid.). This result proved that the spin frequencies of atomic nuclei can take only discrete values.

However, the above findings supported only the general conjecture of Pauli; values of specific spin frequency of atomic nuclei were yet to be measured. The American physicist Isidore Rabi modified the Stern-Gerlach technique in 1933 and was able to measure the spin frequencies of atomic nuclei of hydrogen. Then, in 1937, after some discussions with the Dutch physicist Cornelis Gorter who had done experiments with oscillating magnetic fields, Rabi decided to extend the Stern-Gerlach technique by including in it an oscillating magnetic field, which is a magnetic field whose intensity

changes over time. When Rabi carried out the Stern-Gerlach experiment with the inclusion of an oscillating magnetic field in addition to the static field, he held the frequency of the oscillating field constant and varied the strength of the static field. The results showed that when the strength of the magnetic field approached the spin frequency of atomic nuclei in the beam, the nuclei absorbed energy from the magnetic field. This was the first demonstration of the phenomenon of nuclear magnetic resonance and Rabi was given the Nobel Prize in Physics in 1944 for his work (Huettel et al., 2008).

In order to work, Rabi's experiments and the original Stern-Gerlach technique required gaseous beams from purified gases. After all, these experiments were done to test conjectures and predictions brought about by quantum theory about the spin and magnetic properties of atomic nuclei, so they were experiments done mostly in the service of pure research in physics. In order for magnetic resonance to be used in practice as a measurement technique, the phenomenon of magnetic resonance had to be demonstrated in solid substances. In 1945, Edward Purcell of MIT started doing experiments with his colleagues to do just that. In their experiment, they placed paraffin wax, a chemical compound, into the center of a magnetic field generated by a strong magnet. After some initial experiments that failed to create any resonance effects, they decided to test all possible magnetic field strengths and finally at a near maximum field strength they got a clear resonance effect with the paraffin wax. The reason for their initial failed experiments was their miscalculation of the amount of current necessary to generate the required magnetic field (Purcell et al., 1946).

At around the same time, another group of researchers at Stanford led by Felix Bloch were also investigating magnetic resonance in bulk matter. Their experimental

apparatus was very different from Purcell's. They used a sample of water in a brass box which they placed between the poles of a strong magnet and they manipulated the strength of the magnetic field generated by this magnet. They sent electromagnetic energy from a transmitter coil into the sample and they measured the changes in the energy absorbed by the water sample using a detector coil. The results showed that Bloch and his colleagues also observed magnetic resonance effects in the water sample and they named this phenomenon nuclear induction. The results of the Purcell group at MIT/Harvard had been published in January of 1946 in the journal *Physical Review*. The results of the Bloch group at Stanford were published in the same journal only two weeks after the Purcell results were reported. What Bloch and his colleagues named nuclear induction later came to be known by its present name, nuclear magnetic resonance (NMR). NMR is the basis of all modern magnetic resonance imaging techniques known by the public as MR imaging or MRI for short. All MR scanners apply the essential design of Bloch's apparatus with a strong static magnetic field, a transmitter coil, and a detector coil. Purcell and Bloch were given a joint Nobel Prize in Physics in 1952 for independently demonstrating nuclear magnetic resonance effects in bulk matter. Starting in the 1950's and in the following decades NMR was widely used as a technique for chemical analyses of substances and proved to be a commercially successful method employed in geology and organic chemistry (for a historical analysis of the development of NMR spectroscopy as a domain of knowledge, see Roberts, 2002) . In the next section, we will take a brief look into the development of magnetic resonance imaging in biological settings.

2.3.2: Magnetic Resonance Imaging in Biology and Medicine: The first medical use of NMR was proposed by the American physician Raymond Damadian. In the 1960's several experiments had shown that atomic nuclei in water molecules in biological tissue had different properties with respect to their diffusion and orientation as compared to atomic nuclei in water molecules that were not in biological tissue. One could identify these differences by NMR procedures. Damadian proposed that there may be similar differences between cancerous and non-cancerous cells, and if so, NMR could be used to identify cancerous tissue. This hypothesis was tested by Damadian in 1971 and he found indeed that NMR identified differences between cancerous and non-cancerous tissue (Damadian, 1971). This was the first medical application of the NMR technique. However, Damadian's work was simply using NMR for assessment of tissue samples for any differences and did not create any images of any biological tissue.

In order for NMR to be significantly useful for biology and medicine there had to be a way of employing NMR to create images of biological tissue. The American physicist Paul Lauterbur knew of Damadian's work, and he saw that if a method could be developed to create images using NMR this would be of great use in physics, biology, and medicine. In 1972, Lauterbur proposed that variations over space in the strength of magnetic field would lead to variations in resonant frequencies of protons at different field locations. If this were true, then one could measure the number of protons present in different spatial locations by measuring how much energy was emitted at different frequencies. Lauterbur's suggestion of inducing spatial gradients in the magnetic field was the essential insight for the creation of MR images. Applying this idea and introducing four different spatial gradients, Lauterbur created the very first MR image in

1973. It was the MR image of two water-filled test tubes. He used four spatial gradients, each turned 45 degrees from the previous one, and obtained successive measurements of the two test tubes. Measurements made under each gradient provided different information about the tubes and all the measurements were combined to reconstruct an image of the tubes that revealed their spatial organization (Lauterbur, 1973).

Although Lauterbur's work launched MR imaging, the method he used was too time-consuming for efficient practical use. In addition, the images it yielded were only two dimensional. In 1976, the British physicist Peter Mansfield proposed a much more efficient technique, which later was called echo-planar imaging. In this technique, an electromagnetic pulse was sent from a transmitter coil into the sample and right after that rapidly changing magnetic field gradients were introduced. During this time the MR signal was also being recorded and this yielded a complex signal, which was then reconstructed into an image using Fourier analyses. Echo-planar imaging made it possible to collect images in fractions of a second, instead of minutes with the older techniques. This was *the* major development that established MRI as an effective tool for clinical imaging. It also needs to be mentioned that the reduction in the time it took to obtain an image was of crucial importance for the later introduction of functional MRI as it is a central necessity to have fast imaging in order to be able to measure changes in brain activity. Lauterbur and Mansfield were given a joint Nobel Prize in Physiology and Medicine in 2003 for their work in the development of MRI (Huettel et al., 2008).

The above developments provided the scientific bases for MRI. Yet, making it work effectively introduced serious engineering problems. One problem in the 1970's was the difficulty of creating strong magnetic fields. This problem plagued several early

attempts to build MR scanners for wide use as imaging tools. However, because of the significant potential of MRI, starting in the 1980's several big companies, such as General Electric, Philips, Siemens, and Varian, started their own projects for the development of MRI scanners to be used in clinical settings. The resources that these companies allocated for MRI research led to substantial increases in the power of MR scanners. Standard resistive magnets were replaced by superconducting magnets; these could create stronger and more homogenous magnetic fields. For example, Damadian's early scanner had a magnetic field strength of 0.05 Tesla (Tesla is the unit of measurement of the strength of a magnetic field), while as early as 1982 General Electric had already produced the first commercial MR scanner whose field strength was 1.5 Tesla. By the middle of the first decade of the 21st century, 1.5 Tesla scanners were replaced by new scanners with magnetic field strengths of 3.0 or 4.0 Tesla (ibid.). MRI has become the most commonly used diagnostic imaging method and its widespread use in medicine set the stage for the development of functional MRI.

2.3.3: Blood-Oxygenation-Level-Dependent Contrast and functional MRI: In the previous two subsections, we have seen how the discovery of a physical phenomenon, magnetic resonance of atomic nuclei, started a trajectory of research that eventually led to the development of an effective imaging device for use in medical settings. This was the development of what is today called *structural* or *anatomical* MRI, the technique that makes use of magnetic resonance to provide images of live biological tissue and is used mainly as a diagnostic tool in medicine. For the development of *functional* MRI a separate trajectory of physiological research had to be completed, which would

eventually provide part of the scientific basis for functional MRI.

Energy is required for the brain to function. As Roy and Sherrington showed as early as 1890, the necessary energy is continuously provided to the brain by blood flow. That is, as some region of the brain responds to a stimulus or as it engages in some motor or sensory or other cognitive function, then that part becomes more active and, as a result, it needs more energy. To provide that energy, there is an increase in the cerebral blood flow to that brain region. Glucose and oxygen are the main energy sources of the brain. Oxygen is attached to hemoglobin molecules, which are proteins in the red blood cells whose job it is to transport oxygen from lungs to any part of the body that is in need of oxygen, including the brain. Hemoglobin transports oxygen to different parts of the brain in various amounts depending on which parts of the brain are more active than others. In those active parts of the brain, hemoglobin releases the oxygen for use by neurons or other types of brain cells such as glial cells. Functional MRI makes use of this biological mechanism to detect varying levels of blood flow in the brain. Thus, fMRI is a tool for indirectly measuring the various levels of activity in different parts of the brain in terms of cerebral blood flow. But let us also remember that MRI makes use of magnetic resonance properties of atomic nuclei, so we must understand how these properties are related to cerebral blood flow and the measurement of blood flow by a tool that detects magnetic resonance. The necessary link between MRI and cerebral blood flow was provided at around the time in the 1930's when the physicist Isidore Rabi was developing experimental techniques for demonstrating the magnetic properties of spin and orientation of atomic nuclei conjectured by Wolfgang Pauli (Huettel et al., 2008).

In 1936, the American chemist Linus Pauling and his student Charles Coryell did a series of studies of the molecular structure of hemoglobin. In these studies, they discovered that oxygenated hemoglobin and deoxygenated hemoglobin had opposite magnetic properties. Oxygenated hemoglobin has no unpaired electrons and its magnetic moment is zero. Thus, it is diamagnetic and is repelled from a magnetic field. In contrast, deoxygenated hemoglobin has unpaired electrons and a significant magnetic moment. Thus, deoxygenated hemoglobin is paramagnetic and is attracted to a magnetic field (Pauling & Coryell, 1936).

The physics of magnetic fields says that when an object with magnetic susceptibility is introduced into a magnetic field, this causes inhomogeneities in that magnetic field. Deoxygenation of blood affects magnetic susceptibility, and as a result, we get higher levels of MR signals in areas of the brain where blood is highly oxygenated and lower levels where blood is deoxygenated. In the early 1980's, Thulborn et al. tested this hypothesis in an experiment. In this experiment, they assessed the relationship between the amount of oxygenated hemoglobin in a test tube of blood and magnetization. They found that the decay of magnetization did indeed depend on the proportion of oxygenated hemoglobin in the blood. They also found that this effect was amplified at magnetic fields of higher strength: at field strengths of less than 0.5 Tesla, no large difference between oxygenated and deoxygenated blood was found with respect to magnetization. But at field strengths of 1.5 Tesla, the difference was substantial (Thulborn et al., 1982). This finding later proved crucial for the development of functional MRI using the structural MRI scanners of the 1990's with their high strength magnetic fields.

The results described above provided the scientific basis for the measurement of variations in blood oxygenation levels using MRI. In the late 1980's, Seiji Ogawa, a research scientist at Bell Laboratories, investigated the possibility of utilizing this phenomenon and studying brain physiology using MRI. However, one problem was that standard MRI scanners used the magnetic properties of atomic nuclei of hydrogen. Hydrogen exists in substantial amounts in water throughout the human body and this would render MRI incapable of detecting small changes in metabolic processes that occur constantly in the body. In order to be used to examine physiological processes, MRI had to be able to detect metabolic processes. Ogawa and his colleagues suggested using blood flow as a measure of metabolism, because metabolic processes require oxygen which is provided by oxygenated hemoglobin in the blood (Huettel et al., 2004). The suggestion of blood flow was somewhat obvious if we remember that in 1963 Ingvar, Lassen, and their colleagues (cited in Raichle, 1998) had developed techniques, which used scintillation detectors placed over the head, that enabled researchers to measure regional blood flow in the human brain. With this technique, they showed that blood flow in the brains of normal human subjects showed regional changes in response to changes in functional brain activity.

On the basis of previous findings that showed how blood oxygenation levels caused magnetic field inhomogeneities, Ogawa and his colleagues proposed that manipulations of proportion of blood oxygenation would have an influence on the visibility of blood vessels in MR images. In 1990, they tested this prediction in a series of experiments in which they manipulated the level of blood oxygenation in rodents. Their results showed that indeed the presence of deoxygenated blood made the vessels appear

as thin dark lines in the MR images. They concluded that deoxygenated blood leads to a decrease in the measured MR signal and this causes the appearance of those thin dark lines where there is more deoxygenated blood than oxygenated blood (Ogawa et al., 1990). This was the finding that later came to be called the blood-oxygenation-level-dependent (BOLD) contrast. Ogawa proposed that this contrast may be used to measure changes in functional activity in the brain using MRI; increased functional activity would lead to an increased demand for energy, which in turn would cause increased levels of oxygen consumption and increased proportions of deoxygenated hemoglobin. This would cause magnetic field inhomogeneities, which would be measured by the MRI scanner. This suggestion paved the way for the first functional MRI experiments in the early 1990's. The primary aim of these experiments was not producing new knowledge about brain activity, but rather to replicate well-known findings that had been established by other fields of neuroscience, such as lesion studies in rodents or electroencephalogram studies with humans. The idea was that if studies using fMRI could replicate these known effects, then this would provide evidence that fMRI would not lead researchers astray when it is used to investigate neural phenomena about which we do not know much. Therefore, these early fMRI studies used visual and motor tasks that human subjects were asked to perform as fMRI data were collected (Huettel et al. 2004). Our knowledge of several effects in the visual and motor systems of the brain has been well-established by other subfields of neuroscience. For example, during a finger-tapping task, we would expect fMRI to detect significant activation in the motor cortex. The results of these experiments showed that fMRI procedures did indeed replicate known findings and consequently the stage was set for the expansion of cognitive neuroscience with fMRI as

its main research tool, leading to the explosion of fMRI research in the late 1990's and the first decade of the 21st century.

2.4: Theory-Ladenness of a Useful Kind

The historical accounts of the different trajectories of scientific research that eventually converged in the development of fMRI are rich in their implications for the philosophy of science. Here, I discuss how this history may give us new insights into the question of theory-ladenness of observations in science.

Claims to the effect that experimental data are theory-laden abound in philosophies of science since the widely influential works of Thomas Kuhn (1962/1996) and Norwood Russell Hanson (1958), who famously stated that "seeing is a 'theory-laden' undertaking" (p. 19), meaning that our observations of objects are laden with our prior knowledge of those objects. Although, as Bogen (2010) observes, Kuhn and Hanson discuss some examples of research that feature observations generated by scientific equipment, such as microscopes or telescopes, their arguments are mainly about observation as a perceptual process in humans. Yet, most contemporary scientific observation is done through the use of complicated research instruments, such as electron microscopes, DNA sequencers, or fMRI. Such instruments work on the principles provided by scientific theories, so one could say that contemporary scientific observation is even more theory-laden when compared to earlier times.

In chapter one, we have seen how some philosophers of science express strong skepticism about the epistemic value of fNI findings. One reason for their skepticism was that fNI findings were too theory-laden with modularist theories of human cognition and

that this leads to circularity in the way fNI experiments are designed and findings reported. This criticism is related to a certain understanding of theory-ladenness in philosophy of science. Specifically, this kind of theory-ladenness occurs when the meanings of the terms used to describe experimental findings are strongly influenced by the scientific theory adopted by the researchers who design and carry out the experiments. For example, as Kuhn argued, researchers who adopt different theories of heat, caloric versus mechanical theories, would disagree on the description and evidential import of data from heat experiments. One can also think of the debate on Uttal's criticisms of fNI research in these terms (2001, discussed in chapter one). One of Uttal's major concerns is how easily fNI results can be erroneously taken as support for modularist theories of human cognition, i.e. those theories that accept the view that cognition occurs through the workings of separate modules located in different parts of the brain. Uttal advocates the use of fNI techniques but rejects the modularist interpretation of fNI data. As such, one can construe the debate between Uttal and his opponents, e.g. Bechtel (2002a), as a fundamental disagreement about how fNI results should be interpreted. This question is inherently related to what type of large-scale theory of cognition one adopts, e.g. modularist versus non-modularist theories.

Some philosophers of science use this issue as support for the underdeterminationist conclusion that we cannot learn much from fNI studies. For example, Hardcastle and Stewart (2002) cite the theory-ladenness of fNI results as the central premise for their skepticism about the epistemic value of fNI as a paradigm for cognitive science. They claim that because fNI results are too theory-laden with modularist theories, when we draw inferences about how the brain works from fNI data

we end up falling into circular arguments. Modularist assumptions, they claim, are located in the hard core of fNI; researchers appeal to these assumptions when they design, carry out, and interpret experimental findings. Yet, these assumptions may in fact be false. Thus, Hardcastle and Stewart conclude that fNI results cannot provide any support for modularist theories, because, prior to experiments researchers already assume the existence of specific cognitive modules. As a result, there is an inherent circularity in fNI research. If Hardcastle and Stewart were right, one could also say that observations in fNI are hopelessly theory-laden. Hardcastle and Stewart also state that at this point we simply cannot do any better, because we have no way of studying relationships between the brain and cognitive processes in a framework other than a modularist one. Of course, other authors, for example Uttal (2001) and Bechtel (2002a) disagree with this bleak conclusion. I discuss how we can avoid problems of this kind of theory-ladenness in fNI research at the end of the chapter three, where I show that we can talk fruitfully about fNI data independently of modularist versus non-modularist theories of cognition. Let us now move on to another kind of theory-ladenness that occurs in fNI research that is in fact useful in establishing the reliability of experimental findings.

In fMRI research, the constructs and phenomena that are of interest to cognitive neuroscientists are represented and investigated using the highly complicated workings of the fMRI machine. The workings of fMRI depend on complex and numerous experimental procedures which are based on well-established experimental knowledge coming from physics and physiology. Some think of this fact as a negative feature of fMRI that lowers or limits the reliability of fMRI findings. For example, as I discussed in chapter one, this aspect of fMRI was the major reason for Roskies's skepticism about the

reliability of fMRI findings (2008; 2010). Bogen (2010) has argued that the dependence of fMRI on these procedures, which are based on theories of magnetic resonance and hemodynamics, calls into question the reliability of fMRI as an observational tool, because it is difficult in fMRI to pinpoint what exactly is observed. He states that fMRI is a type of science where "evidence is produced by processes so convoluted that it's hard to decide what, if anything has been observed" (ibid., p. 11).

I propose that the dependence of fMRI on these complex procedures is a useful kind of theory-ladenness. The workings of fMRI depend on well-established knowledge of nuclear magnetic resonance and magnetic characteristics of hemodynamic processes, which come from physics and physiology, respectively. Thus, fMRI findings are laden in the theories of physics and physiology. This is related to a distinction Pierre Duhem (1906/1991) raised between physics and physiology; the physiologists make their observations using measurement techniques based on the established theories of physics, whereas physicists have to test their theories based on the theories of physics. When we look at fMRI, we see that cognitive neuroscientists have to rely on not only physics but also on physiology. Thus, at first look, it may seem that cognitive neuroscientists are at a yet higher level of theory-ladenness than Duhem's physiologists, because we obtain fMRI findings on the bases of theories of both physics and physiology. However, when we remember that these theoretical bases sit on solid foundations provided by experimental knowledge that these sciences have produced, it appears that this kind of theory-ladenness is not necessarily a bad thing, because it allows for the representation of and intervention in cognitive neuroscientific constructs using well-established concepts and methods of physics and physiology.

This is related to Hacking's (1983) discussion of how the reality of dense bodies in blood is established using different types of experimental techniques, namely, electron microscopes and fluorescent staining. Experiments using either technique yield the same result, specifically in both types of experiment small dots in red blood platelets are observed. Hacking appeals to an argument from coincidence about the reality of these findings; it would be a highly improbable coincidence that independent procedures yield the same result unless these small dots are real entities rather than instrumental artifacts. He states that the dense bodies are real entities because experimental instruments using different physical theories yield the same observations and also because we have a clear understanding of the physical theories that are used to build those instruments. Earlier in this chapter, we have seen that the initial fMRI experiments were done on cognitive processes of which we have a clear understanding from previous research in behavioral studies with humans and lesion studies using animal models, e.g. perceptual or motor functions. These initial fMRI experiments were done to check for the reliability of fMRI as an experimental tool and it was seen that fMRI experiments yielded observations that agree with previously established findings. For example, we know from previous non-fMRI research that visual perception is related to activation in a certain area of the brain known as the occipital cortex and fMRI experiments where subjects did visual tasks have shown that they had high degrees of activation in their occipital cortex. Thus, it would be a "preposterous coincidence" that various types of experiments using different tools and paradigms yield the same kind of artifactual result unless visual functions are indeed related to activation in the occipital cortex. These initial experiments helped establish fMRI as a useful tool for cognitive neuroscience. In addition to this, we have a clear

understanding of the theories on which fMRI as an instrument is built. These are physical theories of magnetic resonance and physiological theories of magnetic properties of hemoglobin. Both theories, as discussed earlier, are based on well-established experimental knowledge. Furthermore, and more importantly, because of our clear understanding of magnetic resonance and hemodynamics, we know what types of methodological errors the fMRI machine may produce and we can control for these errors. Therefore, the physical and physiological background theories of fMRI provide more reliable tests of research hypotheses in cognitive neuroscience. (I discuss how this kind of error analysis can be done in chapter four.) All this provides additional support for the reality and manipulability of the constructs of cognitive neuroscience independently of large-scale theories of human cognition. Now, let me elaborate on this notion of useful theory-ladenness by discussing an fMRI study as a concrete example.

In a series of fMRI experiments, John Gabrieli and his colleagues (1996) have studied the neural substrates of semantic memory tasks, namely encoding and repetition priming of meanings of words. The researchers used 480 words, one half of the words were abstract (e.g. trust) and the other half were concrete (e.g. chair), which were randomly grouped into sets of 20 words. The words were displayed in uppercase or lowercase letters; that is, each 20-word set contained 5 abstract words in uppercase, 5 abstract words in lowercase, 5 concrete words in uppercase, and 5 concrete words in lowercase letters. The subjects were asked to perform two cognitive tasks; the semantic encoding task in the first phase of the study and the repetition priming task in the second phase. In the *semantic encoding* task, they were asked to judge whether the word shown on the screen was an abstract or concrete word. As a control for this task, they were also

asked to do a *perceptual encoding* task in which they were asked to judge whether the word shown was in uppercase or lowercase letters. Previous studies have shown that semantic encoding works more efficiently for a repeated word than for the initial presentation of that word, this is known as repetition priming and is an example of memory retrieval. So, to study memory retrieval, Gabrieli and his colleagues have asked their subjects to do a *semantic repetition* task in the second phase of their experiment. The subjects were to do the same semantic encoding task from the first phase, that is making judgments of words being abstract or concrete, but this time every second word set included the same words as the previous set but in a different order. Indeed, the subjects made semantic judgments more quickly for repeated words than for words that were not shown before, so the subjects showed repetition priming. As the subjects performed these tasks, fMRI data were collected.

Results of fMRI scans were compared across tasks. For the encoding task in the first phase of the study, fMRI data were compared between semantic and perceptual encoding using statistical tests (t-tests) and correlation analyses. The results showed that subjects had a significantly higher amount of activation in the left inferior prefrontal cortex when they performed the semantic encoding task compared to the perceptual encoding task. For the semantic repetition task in the second phase of the study, fMRI data were compared between the first and the second presentation of words when subjects were again to make semantic judgments, i.e. abstract or concrete. The fMRI results showed that when they did this task for repeated words there was less activity in the left inferior prefrontal cortex compared to words that were not shown before. On the bases of these results, Gabrieli and his colleagues have concluded that the left inferior prefrontal

cortex is involved in semantic working memory. They reason that this is because more semantic information is needed for semantic encoding than for perceptual encoding, and the results showed higher activation in this region in the semantic encoding task than for the perceptual encoding task. In addition, this region was shown to have less activation for repeated words in the semantic repetition task, which shows that it becomes less active as the demand on semantic working memory decreases. In the case of repeated words, the semantic judgment is made more easily and quickly, because of implicit memory retrieval, and the demand on semantic working memory is reduced. Since less activity was observed in the left inferior prefrontal cortex, this region may be involved primarily with semantic working memory.

Regarding the results of this experiment, one could argue, as Uttal (2001) and Hardcastle and Stewart (2002) certainly would, that the findings of this study are too laden in modularist theories of contemporary cognitive science. The researchers seem to already assume before the experiment that semantic working memory and perceptual working memory exist as separate cognitive modules. Who is to say for certain that hypothetical constructs such as semantic working memory really do exist as modules in the brain? One could raise the same question about the reality of distinct cognitive functions such as semantic encoding or perceptual encoding as defined by theories of cognitive psychology. In fact, I agree that this is a genuine worry, but contrary to Hardcastle and Stewart, I think it is possible to make sense of fMRI findings without making any commitments to modularist theories. A requirement for this is a clear and precise understanding of what fMRI data can reliably tell us. Gaining this understanding is the project of chapter three at the end of which I discuss how we can interpret fMRI

findings without committing to large-scale modularist theories. This would protect us from falling into circular arguments as we discuss fMRI findings but at the same time we can avoid rendering fMRI findings theoretically useless.

What I wish to emphasize here is that the Gabriela et al. study is a series of fMRI experiments. As such, its findings are laden with physical theories of magnetic resonance and physiological theories of magnetic properties of hemodynamic processes in the brain. Any fMRI study has central assumptions, these are: 1) If a certain brain region is involved in the performance of some cognitive task, then when subjects perform that task that region of the brain will be more active; 2) This activity will cause increased blood flow to that region, which will change the concentration of oxygenated and deoxygenated hemoglobin; 3) The altered concentration of oxygenated and deoxygenated hemoglobin will lead to magnetic field inhomogeneities in that region of the brain, which the fMRI scanner will detect as the blood-oxygenation-level-dependent (BOLD) response. All this we know because of the well-tested, well-established experimental knowledge on magnetic resonance and hemodynamics. What we have as fMRI findings of this study are a series of physiological events (cerebral blood flow), which have an effect on magnetic fields and this effect is picked up by the fMRI scanner. The crucial thing to note here is that these events took place because the researchers gave the subjects different cognitive tasks to do and we got different fMRI results. In other words, the researchers were able to represent some cognitive functions in terms of operationally well-defined experimental tasks and succeeded in manipulating brain activation as measured by fMRI. Thus, we have evidence, using well-tested phenomena of magnetic resonance and hemodynamics, for the difference between semantic encoding and perceptual encoding, as operationally

defined in this study and as shown by the different amounts of activation in the left inferior prefrontal cortex. This finding holds regardless of the truth or falsity of large-scale modularist theories of cognition. This finding also lends support for the reality of the difference between the cognitive constructs of semantic and perceptual encoding. Even if they do not exist as separate modules, they exist as different patterns or amounts of brain activation detected by the fMRI scanner. In this sense, fMRI can enable researchers render cognitive constructs concrete in terms of fMRI findings which in turn are defined in terms of well-established causal effects of hemodynamic phenomena on magnetic resonance. Using fMRI, cognitive neuroscientists can *represent* and *manipulate* cognitive processes in terms of the phenomena of hemodynamics and magnetic resonance. Thus, we can say that if we can image them, then they are real. Of course, the argument here assumes that the fMRI experiments were done without any serious methodological flaws and what we have as the finding is not an artifact but a real effect. How can one make sure that fMRI findings constitute real effects? And, if they do, what do they mean? The error-statistical account and the fact that our knowledge of hemodynamics and magnetic resonance is well-tested can help us answer these questions. It is to this project I now turn in chapters three and four.

Chapter Three:

Primary Models in fMRI:
Determining What Is and What Is Not Underdetermined By Data

3.1: Introduction

In the previous chapter, I provided a historical account of the development of fMRI, which came to be the essential research technique of cognitive neuroscience as two independent trajectories of scientific research converged, namely the study of magnetic resonance in physics and the study of hemodynamic processes in the brain in physiology. In addition to providing crucial insights about theory-ladenness in cognitive neuroscience and in turn the characteristics of research in this field, this account also provided explanations of how and on what principles the fMRI scanner works. This knowledge is indispensable for this chapter and the next, in which I address the questions what we can learn from fMRI and how we can learn it.

In chapter one, I reviewed the literature on the epistemological questions about functional neuroimaging (fNI). Since fNI is still a new science, the literature on its epistemology is not vast and I have argued that the works in it can be grouped into two categories: *Category One*: Discussions of the theoretical significance of fNI findings. *Category Two*: Discussions of the methodological difficulties of obtaining reliable inferences from fNI data. The work in category one focuses on the question of what we can learn from fNI results about human cognition and how, if at all, this knowledge can be used in evaluating theories of human cognition. I have also discussed how this kind of work fails to emphasize methodological features of fNI tools, despite the fact that the

discussion of these features is crucial for understanding the theoretical significance of fNI findings. The work in category two emphasizes two features of fNI: 1) The high degree of complexity of the workings of fNI tools and experiments, and 2) The immensely large data sets that fNI studies yield and the difficulty of statistically modeling and obtaining reliable inferences from these data. Unfortunately, some of the work in this category includes common misunderstandings of statistical methods. Consequently, these authors end up giving in to skeptical arguments and conclude that hypotheses of cognitive neuroscience are underdetermined by fNI data and we cannot learn much from fNI experiments save perhaps for some very broad generalizations that can be taken as heuristics for further studies in other fields of psychological science. For example, Roskies (2008; 2010) thinks that the complexity of the workings of fMRI with numerous computational procedures, including preprocessing of raw data and statistical analyses, create this "inferential distance" between raw data and final neuroimages, which is so big that it cannot be characterized sufficiently or accurately. On the basis of this claim, Roskies concludes that we cannot have any complete appraisals of the reliability of fMRI as an experimental paradigm. Klein (2010a; 2010b) thinks that the major problem in fMRI studies is that they are dependent on statistical significance tests. He accepts as true what a certain portion of the literature on significance tests in psychology claims to be the case, namely that significance tests are flawed. Klein looks at some problems that arise in the use of these tests in fMRI studies, such as thresholding or the large N problem (discussed in chapter one), and concludes that, since fMRI is dependent on the use of these tests, it cannot give us any reliable inferences to statistical or substantive

hypotheses. All fMRI can do, according to Klein, is give researchers heuristics for how to proceed with further non-fMRI research.

In chapter one, I also described Mayo's (1996; 2005a;2005b) error-statistical (ES) account emphasizing the notions of severe tests and the hierarchical framework of models of inquiry and argued that this account has the kind of conceptual machinery that can resolve problems such as the ones raised by Roskies and Klein as well as others. The application of the ES account to the epistemological problems in fNI can do a much better job than giving in to a generalized doubt about the epistemic value of fMRI. The starting point is formulating the problems of the methodology of fMRI in ES terms. First, we need to address the question 'what kinds of hypotheses can be put to severe tests in fMRI studies?' In addressing this question, we clarify the evidential import of fMRI data. We can do this by applying the hierarchical framework of models of inquiry to fMRI. In this framework, we break down fMRI experiments into their parts, and we place these parts from hypotheses to be tested to experimental design and statistical models and analyses of data in the primary models, experimental models, and data models. We do this to find out what hypothesis can be tested by the experiment and how, if at all, any parts of the experiment in its various models influence the results and introduce errors. In this process, we find out what kind of hypotheses can be put to severe tests by fMRI. Once we know that, we can assess, on a case by case basis, whether or not the given experiment constitutes a severe error probe of the hypothesis it is meant to test. Thus, we can answer the questions 'what can we learn from fMRI?' and 'how can we learn that which we can learn from fMRI?' In this chapter, I first provide a more detailed discussion of the hierarchical framework of models of inquiry as described by Mayo (1996), then I

answer the question ‘what can we learn from fMRI?’ by finding out what goes into the primary models of a typical fMRI experiment.

3.2: The Hierarchical Framework of Models of Inquiry and Severe Tests

One crucial aspect of experimentally testing scientific hypotheses is the fact that it is not just about the data and the hypotheses; a great number of considerations are involved in coming up with a research hypothesis or question, designing an experiment to collect relevant data, and organizing, modeling, and analyzing the collected data. As Mayo writes, "an adequate account of experimental testing must not begin at the point where data and hypotheses are given, but rather must explicitly incorporate the intermediate theories of data, instruments, and experiment that are required to obtain experimental evidence in the first place" (1996; p.128). To do this, we need to look at an experimental inquiry in terms of a hierarchy of models that connect the primary scientific hypothesis or question being investigated to the "nitty-gritty details of the generation and analysis of data" (ibid.). We can define, for any experimental inquiry, three types of models; models of primary scientific hypotheses, models of experiment, and models of data. These models help us describe clearly the local procedures that are required to establish the connection between raw data and the substantive hypotheses of interest. This account of the hierarchical framework of models provides just the kind of conceptual framework needed to address the problems in fMRI experiments. Let us delve into these models in more detail.

The primary model includes the local or “topical” hypotheses, which may have been derived from a higher-order scientific theory or from previous studies in the field,

and they correspond to a given primary question or problem. This primary problem typically ends up being an evaluation of a quantity associated with a model or theory. In the case of fMRI, the primary hypothesis often takes the form of predicting the amount of brain activity, or the value of signal intensity, in a given brain region as measured by the fMRI scanner. The experimental model provides the link between the data collected in the experiment and the primary hypothesis being tested. Mayo (1996) talks about two functions of the experimental model. The first function is to provide "a kind of experimental analog of the salient features of the primary model" (p. 133). If the primary problem is testing a hypothesis, then the experimental model tells us what is expected to obtain in this experiment if the hypothesis is true, possibly by using other auxiliary hypotheses. The second function of the experimental model is "to specify analytical techniques for linking experimental data to the questions of the experimental model" (p. 134). If the primary problem is testing a hypothesis, because of the many sources of error that influence the data collection process, the data will very rarely, perhaps never, agree exactly with the experimental prediction. In this case, the experimental model may statistically formulate the link between the primary hypothesis and the data model. Thus, the first job of the experimental model is to link the primary model (primary hypotheses) with the experimental model (experimental hypotheses), and its second job is to link the experimental model (experimental hypotheses) with the data model (experimental data).

The data model provides the answers to two types of questions: the 'before-trial' question and the 'after-trial' question. The before-trial question is about how raw data should be collected and modeled to be put in canonical form in order to be linked to the experimental model. The after-trial question is about how we can check whether or not

the data collection procedures were in line with the assumptions of experimental models. Mayo (1996) states that "data models, not raw data, are linked to the questions about the primary theory, and a great deal of work is required to generate the raw data and massage them into the canonical form required to actually apply the analytic tools in the experimental model" (p. 135). In the case of an fMRI experiment, the preprocessing of raw fMRI data, in which processes like signal averaging and spatial and temporal filtering are carried out, would be placed in the data models. The application of the hierarchical framework of models to fMRI experiments can help us delineate and appraise the experimental procedures that are involved. More importantly, we can also assess to what extent, if at all, these procedures influence the results independently of the truth or falsity of hypotheses of interest and thus may introduce errors. Thus, we can assess the reliability of the inferences drawn on the basis of those results. It can even give us the kind of the characterization of inferential distance, or something very close to it, and the distance, as it were, can be trodden safely, something that Roskies thinks can never be done. Let us see how a first-pass at breaking down an fMRI experiment into its parts would go.

We can start by looking at the kinds of hypotheses that fMRI experiments are meant to test. These often are predictions derived from large-scale theories of human cognition or results of previous studies. They often take the form "brain region X is involved in the performance of cognitive task C." From this substantive hypothesis we can derive the prediction that "when subjects perform C, there is going to be a significant amount of activation in region X of their brains." This prediction can be placed in the primary model. I will refer to this kind of prediction as the "real effect hypothesis" as it

predicts the real effect of significantly increased activation in a certain brain region when subjects perform a cognitive task.

Next, we come to the experimental model. As stated above, the first job of the experimental model is to establish the link between the hypothesis in the primary model and experimental analogs of what that hypothesis is talking about. In the case of the fMRI experiment, this is done by stating what would happen in the experiment if the primary hypothesis, namely the real effect hypothesis, is true: "as subjects in the fMRI scanner perform an example of the cognitive task C, the scanner will register a significantly high amount of activation in region X of their brains." This, then, would be placed in the experimental model and it would establish the link between the experimental hypotheses and the data models.

Finally, we come to the data model. In the case of fMRI, designing the experiment, choosing the cognitive tasks, that is, the experimental task, would be placed in data models. The design of the experiment ensures that we get the kind of data that will enable us to test if the primary hypothesis of interest is true or false. Thus, the design of an experiment helps connect the actual experiment to experimental models and primary models of the inquiry as a whole. Of course, once we have raw data, we have to put it in canonical statistical form so that we can carry out statistical analyses to draw inferences. In the case of fMRI, this process is difficult and highly complicated because fMRI scanners yield extremely messy and complex data sets and the raw data have to be preprocessed to be rendered ready for statistical analyses. So, decisions about preprocessing steps, such as the use of spatial and/or temporal filters, signal averaging, and any other necessary procedures for preparing raw data for analyses would be placed

in data models. Once we have fMRI data that are ready for statistical analyses, we have to model the data to obtain a statistically adequate model of the data generating mechanism and then carry out significance tests to make inferences about the primary hypothesis. All these procedures would also be placed in the data models. The figure below (fig. 3.1) shows how an fMRI experiment would look when it is broken down to its parts in the hierarchical framework of models of inquiry.

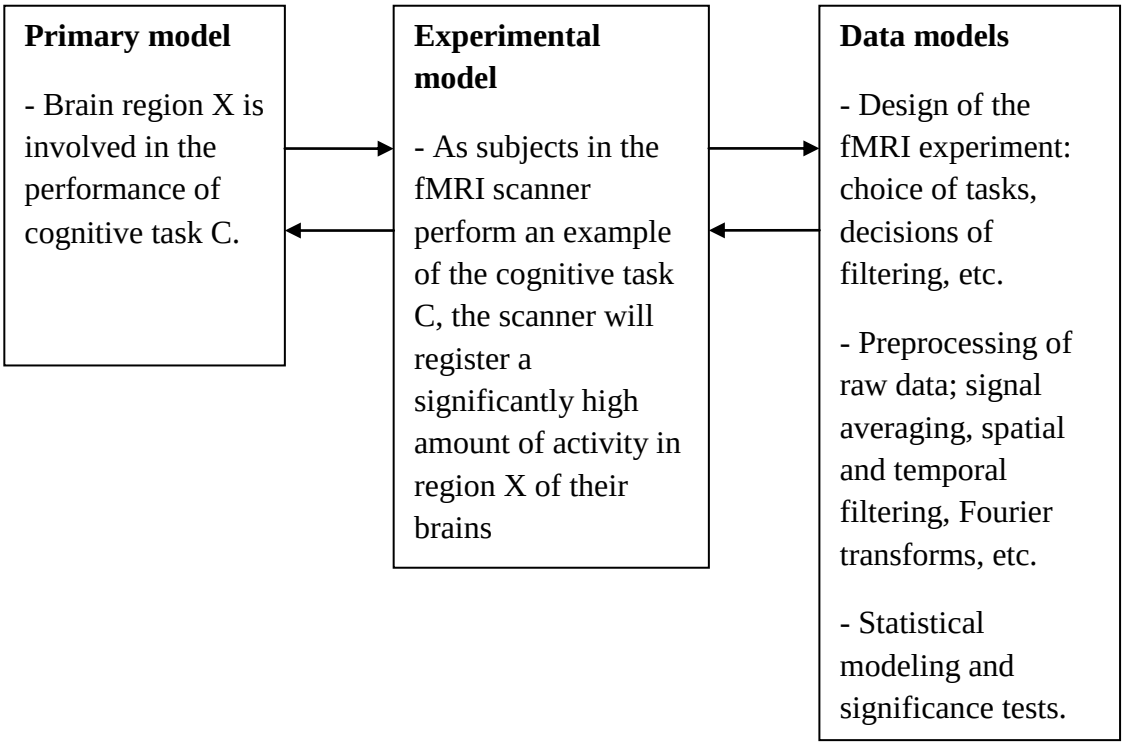

Figure 3.1: Models of Inquiry in an fMRI Experiment: "First Pass"

Although the above figure provides a summary sample of the kind of analysis I wish to carry out, the first thing to note about this figure is that it does not come near to doing justice to the highly complex structure of fMRI experiments. Indeed, Mayo has

written; “Precisely how to break down a given experimental inquiry into these models is not a cut and dried affair—most inquiries will call for several of each” (1996; p. 129). Any given fMRI experiment will certainly require several of these models. The figure above illustrates how we can start with the application of the hierarchical framework of models in dividing up the landscape of an fMRI experiment. This framework fulfills, as Mayo writes, our need to “have at our disposal a framework that permits us to delineate the relatively complex steps from raw data to scientific hypotheses, and to systematically pose the questions that arise at each step” (ibid.). So, in the rest of this chapter and the next chapter we will discuss in detail the primary models, experimental models, and data models of a typical fMRI experiment. As we progress, we will answer the question of what goes into each model and in the process we will have a much more accurate and more complete understanding of how an fMRI experiment works and what it can and cannot give us. Before we go on, let us note a few crucial points about severe tests, which will guide our path through the models of inquiry.

The major point of applying the framework of models to experiments is to answer questions of underdetermination and evidential import of data. In general, these are questions such as ‘what do the data we get really mean in the context of different scientific hypotheses and theories?’ ‘For what specific hypothesis can we take a given data set as evidence?’ These questions turn on another general question; i.e., ‘when do we have evidence for a hypothesis that we want to test in an experiment?’ We can answer this question by taking Mayo’s severity principle as a guide, which states: "Data **x** (produced by process *G*) provide a good indication or evidence for hypothesis *H* (just) to the extent that test *T* severely passes *H* with **x**" (Mayo, 2005a; p.100). As discussed in

chapter one, for a hypothesis to pass a severe test *T* with **x**, two things must obtain; *first*, data **x** fits or agrees with *H*, and *second*, test *T* would have produced, with high probability, data that fit less well with *H* than **x** does, were *H* false (Mayo, 1996; Mayo, 2005a). The idea here is that data **x** is evidence for hypothesis *H* just to the extent that the accordance between **x** and *H* would be difficult to achieve were *H* false. In other words, one must have done a good job at probing the ways one may be wrong in inferring from an accordance between data **x** and hypothesis *H* to an inference to *H* (as well tested or corroborated). The severity of a test is not a feature of only the test itself. It is a function of three things; namely, the *test*, or the experiment; the *data*; and the specific *hypothesis* about which an inference is drawn (Mayo, 2005a). So, we can use the abbreviation SEV(*T*, $\mathbf{x}_0$, *H*) to mean "the severity with which *H* passes test *T* with data $\mathbf{x}_0$" (Mayo & Spanos, 2006; 2010). This abbreviation expresses the severity function SEV(*T*, $\mathbf{x}_0$, *H*) which is assigned a quantitative value between 0 and 1 when the necessary calculations are made; the closer the value to 1 the more severe is the test.[3] As a matter of fact, the project of this chapter and the next can be thought of as filling in the arguments for the severity function.

Before moving on, it should be noted here that although the above is mainly about experiments and statistical tests, the notion of severity can also be employed in discussing error characteristics or error probabilities of specific experimental instruments. In any given experiment, we need to know the error characteristics associated with the components of that experiment, such as experimental design, instruments used for measurement and data collection, processing of data and statistical modeling and analyses

[3] In some cases, severity evaluations can be qualitative.

in order to assess whether or not the experiment is free of errors or possible errors are well-controlled for.

In the context of fMRI, the question of evidence becomes ‘when do we have evidence or support for a hypothesis that we want to test in an fMRI experiment?’ If, as stated above, the severity of a test or experiment is a function of three things, namely, the *experiment*; the *data*; and the specific *hypothesis*, then in order to answer the question of evidential import of fMRI results, we have to have accurate characterizations of each of these three aspects of an inquiry. That is, we have to know what hypothesis the researchers want to test, how the experiment is carried out to generate data, and what the data look like. We can learn all this by breaking down an fMRI study into its primary models, experimental models, and data models. Once we place components of an fMRI study in their proper places in the framework, we can, and indeed must, ask 'what are the error probabilities associated with these components?' Do they introduce any errors? The first thing to identify is the specific hypothesis that researchers want to test in an experiment, so we begin by asking ‘what goes into the primary models of a typical fMRI experiment?’ Or, in other words, ‘what do we take as the hypothesis H in the severity function SEV(T, $\mathbf{x}_0$, H)?’ In the rest of this chapter, we will answer this question, and in the next chapter we will cover the experimental models and data models. At the end, we will have a framework in which each type of model in the hierarchy is filled in with their respective components. This framework will help us formulate the questions of what kinds of hypotheses fMRI data can provide evidence for and how we can assess fMRI experiments regarding their severity in testing these hypotheses. The answers to these

questions will help us resolve questions of underdetermination and also teach us some general lessons about the nature of inference in cognitive neuroscience.

3.3: Primary Models in fMRI

Functional magnetic resonance imaging (fMRI) is the most common neuroimaging technique in cognitive neuroscience. One major goal in an fMRI experiment is to relate changes in brain physiology over time to an experimental manipulation; e.g., looking at a picture of an apple or reading the word 'apple'. This relation is established in terms of the hemodynamic response, also known as the blood-oxygenation-level-dependent (BOLD) response. In a typical fMRI experiment, there are a control group and an experimental group, or several experimental groups, and the experimenters check for statistically significant differences between levels of brain activation across these groups. The fMRI scanner measures changes in the blood-oxygenation-level-dependent (BOLD) response, which occur as a product of hemodynamic processes in the brain. The basic idea behind fMRI is that if a certain brain region is involved in the performance of a given cognitive task, then when an individual performs that task there is going to be increased activation in that region of the brain. Increased activation causes an increased need for energy and this leads to an increase in local glucose metabolism and oxygen consumption. Oxygen is carried to cells by oxygenated hemoglobin in the blood. Oxygenated hemoglobin and deoxygenated hemoglobin have different magnetic properties. In activated areas of the brain there results an imbalance between concentrations of oxygenated and deoxygenated hemoglobin and this leads to inhomogeneities in the magnetic field. The fMRI scanner detects these inhomogeneities and thus provides data on hemodynamic activity in the

brain in terms of the magnitude of the BOLD response (Huettel et al., 2004). It is assumed that the higher the activation in a brain region, the higher is the magnitude of the BOLD response observed in that region. Of course, the discussion of how fMRI works as an experimental paradigm is much more detailed than is discussed here involving complicated physics of spin and magnetic fields some of which will be discussed in the next chapter when we delineate the experimental and data models. However, this brief summary of how fMRI works is necessary here to make the points about what goes into the primary model of an fMRI experiment. The key thing is that the fMRI scanner measures and collects data on hemodynamic activity in the brain.

Huettel et al. (2004) state that in a typical fMRI study three distinct kinds of hypotheses are involved: *1) Hemodynamic Hypothesis*: This is a hypothesis about hemodynamic activity in the brain and it is the most specific one of the three kinds of hypotheses, because hemodynamic activity is what is measured by the fMRI scanner. In most fMRI experiments, this hypothesis predicts significantly higher hemodynamic activation in the experimental condition when subjects perform cognitive tasks than the control condition. *2) Neuronal Hypothesis*: This is a hypothesis about neuronal activity. Since fMRI does not directly measure neuronal activity, we cannot know the amount of neuronal activity that took place in an fMRI experiment. Nonetheless, some researchers assume that if the observed hemodynamic activity is high then neuronal activity is also high. Of course, this assumption is questionable. *3) Theoretical Hypothesis:* This is essentially a hypothesis about cognitive function. Researchers often use fMRI results to address questions about how cognitive processes work, such as memory, perception, attention, and the like, including how they are realized by which neural structures and

processes. So, the theoretical hypothesis may simply be a proposition that some specific structure or area in the brain is involved in the realization of a specific cognitive function. Thus, at first look, these three kinds of hypotheses are to be placed in the primary models of an fMRI experiment (see figure 3.2).

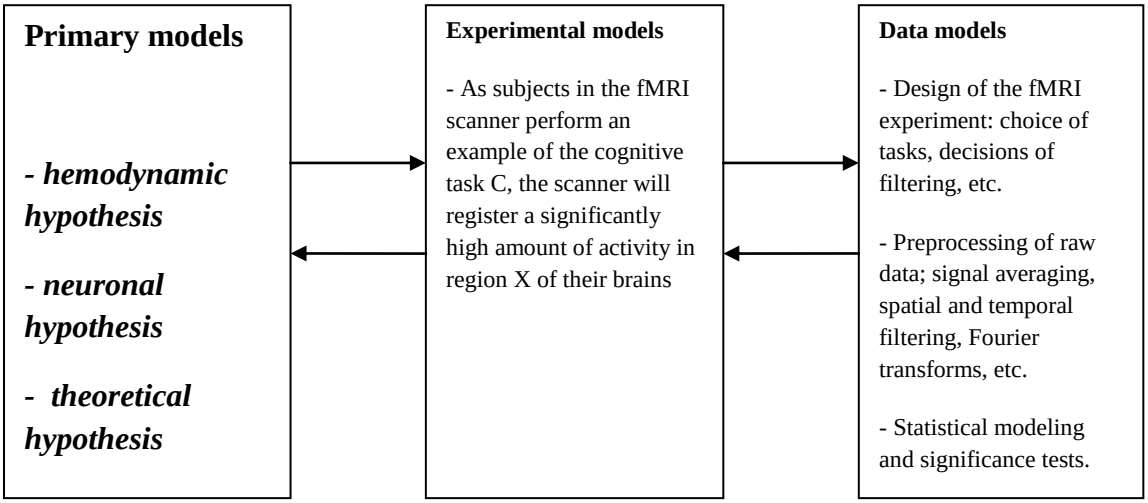

Figure 3.2: Models of Inquiry in an fMRI Experiment: "Second Pass"

It may be said that the central motivation for the establishment of cognitive neuroscience as a research field, with fMRI as one of its central tools, was to finally have the ability to test the theoretical hypotheses of the kind that proposes certain neural substrates for certain cognitive functions as discussed above. However, whether or not fMRI experiments can put all these three hypotheses to severe tests is another question, and indeed, in the methodological approach taken here, it is the central question: which of these three kinds of hypotheses can we put to severe tests in an fMRI experiment? Indeed, it may be asked if these different kinds of hypotheses each require distinct

experimental models. Since fMRI does not give us data on anything other than hemodynamic activity, it appears that hypotheses about hemodynamic activity are the ones we can, at least potentially, put to severe tests in fMRI experiments. Hypotheses about hemodynamic activity in the brain can be formulated in terms of statistical hypotheses that can be tested against fMRI data. Statistical hypothesis testing is widely used in fMRI research, so a brief review of how these tests work is useful.

3.3.1: Statistical Hypothesis Testing and Error Probabilities: Hypothesis testing always occurs in the context of a well-defined statistical model, *M*. The statistical model is a set of independently testable probabilistic assumptions which provide "an approximate (and idealized) representation of the process generating data $\mathbf{x}_0 := (x_1, x_2, \ldots x_n)$" (Mayo & Spanos, 2011; p. 2). The statistical model is defined in terms of the probability distribution, $f(\mathbf{x};\theta)$, of the sample $\mathbf{X} := (X_1, \ldots X_n)$. The θ is an unknown parameter governing the probability distribution of the sample. In statistical analyses, we are interested in certain events that belong to the sample space and the statistical model deductively assigns probabilities to the events of interest in the sample space. But a major purpose of carrying out statistical analyses is to make inductive inferences about the unknown parameters from the sample. This is done when we make inferences from data to statistical hypotheses. The statistical hypotheses are defined in terms of the unknown parameter θ governing the probability distribution, $f(\mathbf{x};\theta)$, of the sample. Thus, statistical inferences are statements about the value of an unknown parameter that is of interest to researchers. For example, one may infer from the data that $\theta > 0$. It is important to note that in error-statistics (ES), probability is *not* used as a degree of confirmation or belief in

inferences like $\theta > 0$. When it comes to drawing inferences from data, probability is used "to quantify how frequently methods are capable of discriminating between alternative hypotheses and how reliably they facilitate the detection of error" (ibid., p. 3). These probabilities that attach to inductive inferential tools are *error probabilities* (which are of central importance in the error-statistical scrutiny of methodological procedures in fMRI that I discuss in chapter four). Statistical hypotheses such as $\theta > 0$ are either true or false regarding the data generating mechanism. The error probabilities are used in statistical significance tests to find out whether or not the data $\mathbf{x}_0$ provide evidence for a statistical hypothesis such as $\theta > 0$.

Statistical significance tests are done in order to make inductive inferences about unknown parameters (θ) of interest. This is made possible by linking data to statistical hypotheses about parameters in terms of test statistics, $d(\mathbf{X})$. Test statistics are functions of the data $\mathbf{x}_0$ and statistical tests are expressed as inference rules defined in terms of the designated test statistic. For example, the rule may be that whenever the value of $d(\mathbf{X})$ is greater than some predesignated constant c, we infer $\theta > 0$. This is a rule for inductive inferences about the unknown parameter θ. As such, and since we have finite data on which to base our inference, we may make erroneous inferences about θ as we apply this test rule. One needs to know how often researchers may end up making erroneous inferences. We can find out the probability of erroneous inferences by calculating the probability of the event $d(\mathbf{X}) > c$ under the assumption that $\theta = 0$. From this we can calculate the probability of erroneous inferences to the hypothesis $\theta > 0$. These error probabilities are calculated from the distribution of the test statistic $d(\mathbf{X})$ which is also called the sampling distribution. Once we have data $\mathbf{x}_0$, we can calculate the observed

significance level, associated with $d(\mathbf{x}_0)$, which gives "the probability of a worse fit with H_0 than the observed $d(\mathbf{x}_0)$, under the assumption that H_0 is true" (ibid., p.4). The significance level or the p-value is a probability: $p(\mathbf{x}_0) = P(d(\mathbf{X}) > d(\mathbf{x}_0); H_0)$. The larger the test statistic the smaller is the p-value. The sampling distribution can be evaluated with respect to different hypothesized values of the parameter θ. One reason that makes error-statistics a very useful framework for fMRI is the fact that it can give us calculations of error probabilities for inferential tools which are numerous and complicated in fMRI studies.

In a typical statistical significance test, we test how well the data $\mathbf{x}_0$ accord with certain statistical hypotheses about the unknown parameter θ. For example, let us look at the case where we have a random sample **X** of size *n* normally distributed with unknown mean μ and known variance σ^2. In this case, we can test the null hypothesis H_0: $\mu = \mu_0$ versus the alternative hypothesis H_1: $\mu > \mu_0$. This is a one-sided test, a type of test used in most fMRI studies, and H_0 and H_1 are defined in terms of the unknown parameter μ. We can test these hypotheses by looking at the test statistic $d(\mathbf{X}) = (\overline{X} - \mu_0) / \sigma_x$ where $\overline{X}$ is the sample mean ($\overline{X} = 1/n \sum_{k=1}^{n} X_k$) and $\sigma_x = \sigma/\sqrt{n}$) (ibid.). The test statistic $d(\mathbf{X})$ is a function of the data $\mathbf{x}_0$ and, given a particular data set, we can calculate a particular $d(\mathbf{x}_0)$. Here, the inductive inference is going from the particular outcome $d(\mathbf{x}_0)$ to a hypothesis about the unknown parameter μ. The test rule is defined in terms of the test statistic: whenever $d(\mathbf{x}_0)$ is greater than 1.96, infer H_1: $\mu > \mu_0$. We can calculate the probability of the event of $d(\mathbf{X})$ exceeding 1.96 under H_0 which gives us the statistical significance level .025. Thus, whenever the observed p-value is smaller than .025 we conclude that there is

evidence of a genuine discrepancy from the value of $d(\mathbf{x}_0)$ expected under the null hypothesis and we infer H_1: $\mu > \mu_0$.

The probability of the event of $d(\mathbf{X})$ exceeding 1.96 under H_0 also gives us the type I error associated with this test rule which is the probability of rejecting H_0 when it is true. Thus, when we fix the significance level at .025, we ensure that the probability of committing a type I error using this test rule is not higher than 2.5 per cent. Let us assume that we do a significance test and get the outcome $d(\mathbf{x}_0)$ is greater than 1.96, so the observed p-value is smaller than .025 and we have a result significant at .025. On the basis of this result we reject the null and accept the alternative hypothesis H_1: $\mu > \mu_0$. Now, we can calculate the severity with which H_1 passes this test with data $\mathbf{x}_0$. By definition of severity, SEV(test *T*, $\mathbf{x}_0$, H_1) = P(test T would *not* yield a passing result; H_1 is false) = 1 – P(test T would yield a passing result; H_1 is false) = 1 - .025 = .975.

We may also commit a type II error which is failing to reject H_0 when in fact H_1 is true. This is given in the probability β = P(Accept H_0; H_1 is true) and to calculate β we have to consider particular point values of μ in H_1, e.g. H_1: $\mu > \mu_1$, where $\mu_1 = \mu_0 + \gamma$ for a positive value of γ. The ability of a test to reject the null hypothesis when the alternative hypothesis is true is known as the power of the test and is given by the probability P(Reject H_0; H_1 is true) which is equal to $1 - \beta$. One major goal in hypothesis testing is maximizing the power of the test and power is always defined pre-data in terms of a rejection rule at a predesignated threshold for significance, say α = .025. The cutoff point c_α at significance level .025 is 1.96. All outcomes beyond c_α = 1.96 are taken as significant; so whenever we obtain a result $d(\mathbf{x}_0)$ which is greater than 1.96, the null hypothesis is rejected and the alternative hypothesis $\mu > \mu_1$ is accepted. In contrast to

power, severity assessments are always carried out post-data with respect to a specific inference. Let us say that a particular outcome $d(\mathbf{x}_0)$ is found to be greater than 1.96, that is, $d(\mathbf{x}_0) > c_\alpha$. Thus, the observed p-value is smaller than.025, so we obtain a result significant at .025. In this case, the severity with which H_1: $\mu > \mu_1$ passes this test with $\mathbf{x}_0$ is given by $P(d(\mathbf{X}) \leq c_\alpha ; \mu=\mu_1)$ which is equal to $1 - POW(T_\alpha;\mu_1)$. Power analyses cannot distinguish between a result that is just beyond c_α as opposed to a result where $d(\mathbf{x}_0)$ is farther away from c_α because both results are treated the same, namely taken as significant results and the null is rejected. In contrast, severity assessments do distinguish between these results; the result that is farther away from c_α yields a more severe test of the same inference $\mu > \mu_1$. This post-data look at experimental results in cases where the null is rejected is very useful, because in tests with high power we may commit, with higher probability, the error of rejecting the null when the outcome is barely beyond the significance threshold. Mayo and Spanos write; "The higher the power of the test to detect discrepancy γ, the lower the severity associated with the inference: $\mu > (\mu_0 + \gamma)$ when the test rejects H_0" (2011; p.21). Since most inferences of interest to fMRI researchers are made in one-sided tests, the difference between power analyses and severity assessments is very useful in determining whether or not we have genuine evidence for alternative hypotheses that state increased hemodynamic activation in certain brain regions. Let us now see how we can formulate statistical significance tests in typical fMRI experiments.

3.3.2: Significance Tests in fMRI: Functional neuroimaging experiments normally have a control condition, where subjects do nothing or do a simple cognitive task, and an

experimental condition where subjects do the cognitive task of interest. Researchers want to see what differences there are, if any, between amounts or patterns of activation in certain regions of the brain across control and experimental conditions. Mostly, the hemodynamic hypothesis, or the real effect hypothesis, predicts that there will be more activation in the experimental condition than in the control condition. We can use μ_0 to designate mean hemodynamic activation (in a certain brain region) in the control condition of an fMRI experiment and we can use μ_1 to designate mean hemodynamic activation (in the same region) in the experimental condition. These two means, μ_0 and μ_1, as well as the difference between them, $\mu_1 - \mu_0$, are unknown parameters, and researchers are interested in making inferences about the difference $\mu_1 - \mu_0$. Statistical hypotheses about the difference between μ_0 and μ_1 can be stated as null and alternative hypotheses in a significance test. We can test the null hypothesis, H_0: $\mu_1 - \mu_0 = 0$ versus the alternative hypothesis, H_1: $\mu_1 - \mu_0 > 0$ in significance tests which are formulated in the context of a statistical model of fMRI data.

The fMRI experiment gives us data on the observed mean hemodynamic activation in every region of the brain in the control and experimental conditions. Once we have fMRI data, $\mathbf{x}_0$, we can use $\overline{X}_0$ to designate the mean of observed hemodynamic activation (in a certain brain region) in the control condition and $\overline{X}_1$ to designate the mean of observed hemodynamic activation (in the same region) in the experimental condition. We can also calculate observed sample standard deviations s_0 and s_1 of the data on hemodynamic activation observed in the control and experimental conditions, respectively. The observed standard deviations s_0 and s_1 can be used as estimates of population standard deviations σ_0 and σ_1. What we need now is a test statistic which will

enable us to test statistical hypotheses about the difference $\mu_1 - \mu_0$. Once we have fMRI data on hemodynamic activation the data on the difference between means can be statistically modeled. The standard deviation of the difference of means $\overline{X}_1 - \overline{X}_0$ is given by $\sigma_{diff.} = \sqrt{(\sigma_1/n + \sigma_0/n)}$ where σ_1 and σ_0 are the standard deviations in the data from the experimental and control conditions, respectively, and n is the sample size which is the same in both conditions. Thus, the test statistic can be defined: $D = \overline{X}_1 - \overline{X}_0 / \sigma_{diff.}$ and we can assign probabilities to outcomes of interest defined in terms of the test statistic *D*. The sampling distribution of *D* under hypothesized values of $\mu_1 - \mu_0$ can be calculated and, given particular outcomes $\overline{X}_0$ and $\overline{X}_1$, we can calculate $D(\mathbf{x}_0)$. We can now define our test rule: whenever $D(\mathbf{x}_0)$ exceeds a predefined cutoff point c_α at $\alpha = .025$ we reject the null hypothesis and we infer H_1: $\mu_1 - \mu_0 > 0$. We set our alpha at .025 and the test rule can now be stated in terms the standard deviation $\sigma_{diff.}$ of the difference of means: whenever $D(\mathbf{x}_0)$ exceeds 0 by 1.96 $\sigma_{diff.}$, infer H_1. Using this test rule and in cases where we infer H_1, we make an inference about the hemodynamic hypothesis of interest which is of the form when subjects perform cognitive task C, there is significantly more activation in brain region X. This is how hemodynamic hypotheses can be tested against fMRI data. In the context of the statistical model of fMRI data, we can calculate error probabilities associated with the test rule. In addition, we can scrutinize the error characteristics of the several different procedures of the fMRI experiment which yield the data. Using these error probabilities we can check, on a case by case basis, whether or not fMRI experiments constitute severe tests of the hypotheses of interest. For example, in an experiment where $D(\mathbf{x}_0) > c_\alpha$ and we infer H_1: $\mu_1 - \mu_0 > 0$, we can check how probable this outcome was if the null hypothesis H_0: $\mu_1 - \mu_0 = 0$ were true using the error

probabilities and the specific outcome $D(\mathbf{x}_0)$. If this probability was high, that is, with high probability the test would have yielded $D(\mathbf{x}_0) > c_\alpha$ even if H_0 were true, then

H_1: $\mu_1 - \mu_0 > 0$ does not pass this test severely with $\mathbf{x}_0$. Severity assessments are always done in terms of a specific inference, H_1, test *T*, and data $\mathbf{x}_0$. Of course, carrying out severity assessments in any specific experimental context is difficult given the complexity of the fMRI procedure, because there may be errors in several different parts of the experiment, from data collection to statistical analysis, which may impair the reliability of the data and lead to fallacies of rejection or acceptance. Questions about such errors will be addressed in detail in chapter four when we look into experimental and data models where I discuss how these errors can be tested and controlled for using the notion of error characteristics. The important thing is that in the context of fMRI experiments, we can test whether or not specific experiments constitute severe tests of inferences to the specific hemodynamic hypotheses tested. Thus, in principle, hemodynamic hypotheses *can* be put to severe error probes in fMRI experiments. This shows that fMRI data do have evidential import for evaluations of hemodynamic hypotheses.

What about the evidential import of fMRI data for the other two kinds of hypotheses, i.e. hypotheses about neuronal activity and theoretical hypotheses about cognitive processes? The neuronal and theoretical hypotheses are not subjected to severe error probes in fMRI experiments, because the fMRI procedure cannot test for the ways in which these hypotheses could be false. The neuronal hypothesis makes an assertion about neuronal activity and we cannot reliably infer much about neuronal activity from the fMRI data on hemodynamic activity. For the observed hemodynamic activity in an

experiment could be due to factors other than neuronal activity, such as activity of glial cells, and in an fMRI experiment we simply cannot test for this possibility (Huettel et al., 2004). In addition, since fMRI does not give us any data on neuronal activation, we cannot calculate any test statistics about neuronal activation. The hypothesis about neuronal activity in a particular fMRI experiment could be false and yet the fMRI experiment would not detect this, because there would be no data against which we can test the neuronal hypothesis. We cannot carry out significance tests of statistical hypotheses about neuronal activity, let alone checking whether or not a test constitutes a severe error probe of a specific neuronal hypothesis. In other words, the test T and data $\mathbf{x}_0$ in the severity function, SEV(T, $\mathbf{x}_0$, neuronal hypothesis), is simply missing from fMRI experiments. If we had evidence from other kinds of experiments that proves a connection between hemodynamic and neuronal activity, then we could establish a link on the basis of such evidence and use fMRI data to evaluate neuronal hypotheses. However, as Huettel et al. (2004) write, though measures of neuronal activity complement fMRI data, "the direct relation between neuronal activity and the BOLD response remains unknown" (p.458). Logothetis (2008), in his discussion of the limitations in the interpretation of fMRI data, reinforces the point that observed hemodynamic activation does not always mean that it was caused by neuronal activity. Logothetis adds that these limitations of fMRI "are unlikely to be resolved by increasing the sophistication and power of [fMRI] scanners" (ibid., p. 876). But perhaps, as some would argue, fMRI can be studied as a research tool by other methods to shed light on the relationship between hemodynamic activity and neuronal activity.

Recent optogenetic studies are promising in this respect. In these studies, populations of genetically targeted neurons are made photosensitive by opsin proteins and this enables researchers to induce action potentials in living animals by stimulation of light-gated ion channels in these neurons. Lee at al. (2010) applied this method to study the relationship between firings of action potentials in neurons and fMRI data: brain activity in rodents was monitored by fMRI while the photosensitive neurons were stimulated by pulses of light, which induced action potentials in these neurons. The results showed that when light pulses were delivered, increased hemodynamic responses were observed in fMRI recordings, which suggests that neuronal activity is indeed one of the causes of hemodynamic activity. Optogenetic studies of fMRI can potentially shed light on the relationship between hemodynamic activity and neuronal activity, but one could question the generalizability of these and other similar findings to human studies. In addition, although optogenetic studies show that increased neuronal activity causes increased hemodynamic responses, this still is just one among several causes, such as activity of glial cells, neuromodulation, etc., that may trigger hemodynamic activity that fMRI picks up as the BOLD response. For example, when we observe increased hemodynamic activity in fMRI experiments with humans, we would still not know if it was triggered by increased neuronal activity or by increased activity of glial cells. Hemodynamic activity may have been caused by neuronal activity related to cognitive tasks, or it may have been caused by activity of glial cells, supporting or protecting brain cells. Thus, the relationship between observed hemodynamic activity in human studies and neuronal activity remains unclear.

Figure 3.3 represents the conclusion of the above discussion. The link between the hemodynamic hypothesis and H_1: $\mu_1 - \mu_0 > 0$ shows that in an fMRI experiment the hemodynamic hypothesis, or the real effect hypothesis, is embedded into a statistical model of the data, it is framed in terms of the parameters μ_0 and μ_1. This connection assumes that the statistical model of fMRI data is statistically adequate. This assumption can be tested for as we will see in chapter four.

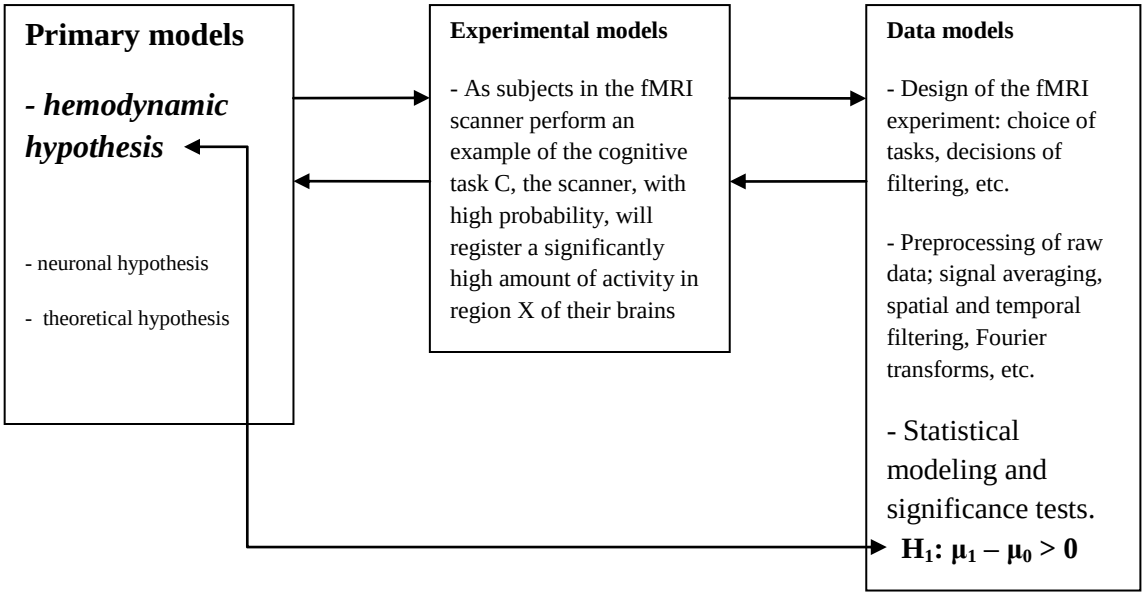

Figure 3.3: Models of Inquiry: Primary Model As Can Be Tested By fMRI

Let me now illustrate the above analysis of identifying what goes into the primary models of an fMRI study in relation to a specific experiment.

3.3.3: Recognition Memory in the Hippocampus: In one fMRI study, Stark and Squire (2000) investigated brain activation patterns in a recognition memory task. Neuropsychological data, animal experiments, and neuroimaging studies have suggested the involvement of structures in the medial temporal lobe of the brain (MTL) for retrieval

and recognition memory. In this study, the researchers sought to study the role of a specific region in the MTL in recognition memory, namely the hippocampal region. First, half the subjects were shown line drawings of objects displayed on a screen (this was the study phase). Then, in the test phase, the subjects were again shown drawings of objects; but this time some drawings were previously shown (targets) and some drawings were of objects not previously shown (foils). The subjects were asked to indicate whether or not the item displayed on the screen was one of the items they had seen before. The procedures were the same for the other half of the subjects except that instead of drawings they studied and were tested on words that named objects. Whether or not there is activation in the hippocampal region in these tasks, which require the formation and use of associations between the presentations of the same stimulus at different times, is of significance for an understanding of the role the hippocampus plays in learning and memory. It has been hypothesized that the hippocampal region is essential for forming and using associations between drawings and words while the adjacent areas of the cortex can handle simpler recognition tasks that do not require such associations. In this study, the researchers sought to test whether or not there would be activation in the hippocampal region across different conditions of the experiment, that is, targets versus foils. The statistical hypotheses that were tested were H_0: $\mu_1 - \mu_0 = 0$ versus H_1: $\mu_1 - \mu_0 > 0$ where μ_0 is the mean hemodynamic activation in the hippocampus when subjects respond to foils and μ_1 is mean hemodynamic activation in the hippocampus when subjects respond to targets. Hemodynamic activation is measured as a continuous variable reporting the amount of activation every two seconds in regions of the brain across the duration of the experiment. In this experiment, $\overline{X}_0$ is the sample mean of observed hemodynamic

activation in the hippocampus across trials when subjects respond to foils, and $\overline{X}_1$ is the sample mean of observed hemodynamic activation in the hippocampus across trials when subjects respond to targets. Thus, the test statistic is $D = \overline{X}_1 - \overline{X}_0 / \sigma_{\text{diff.}}$. The results of two experiments conducted at two different fMRI facilities showed that there were significantly high amounts of hippocampal activation when subjects responded to targets, i.e. words shown at both study and test, that is, in both experiments $D(\mathbf{x}_0) > c_\alpha$. High hippocampal activation was also observed when the other half of the subjects responded to names of objects shown at both study and test phases. So, what exactly do the results of this study show about brain activity? About the hippocampal region? About recognition memory and retrieval?

Since this is an fMRI experiment, we can analyze it in terms of its primary models, i.e. the three different kinds of hypotheses discussed above. First, what is the hemodynamic hypothesis in this experiment? This is a hypothesis about a change in the fMRI signal in a certain region of the brain when subjects perform the experimental task. The experimental manipulation in this study is the presentation of objects, which were either shown before in the study phase (targets) or objects that were not shown before (foils). (Both words, naming objects, and drawings of objects were used as stimuli.) As discussed above, previous research suggests that the hippocampal region is going to be activated when targets are shown because of the associations supposed to be formed between presentations of the same stimulus at different times. Thus, the prediction that was tested in this experiment is this: "there is going to be a significantly higher level of *hemodynamic activity* in the hippocampal region when targets are shown in comparison to when foils are shown." In canonical form, the alternative hypothesis to be tested was

H_1: $\mu_1 - \mu_0 > 0$ and this alternative hypothesis is nothing other than the hemodynamic hypothesis of interest framed in terms of statistical parameters. Thus, the hemodynamic hypothesis is the general assertion that when subjects respond to targets, i.e. stimuli they were shown before, there is an increased amount of hemodynamic activation in the hippocampal region. The results of the experiment showed that indeed there was a significant difference between the target and foil conditions and the researchers inferred H_1: $\mu_1 - \mu_0 > 0$ which is directly related to the hemodynamic hypothesis of interest. Assuming that the experiment was done without serious flaws and the statistical model of the data was statistically adequate, then the results of this study provides support for this hemodynamic hypothesis.

What about the neuronal hypothesis? This is a hypothesis about *neuronal activity* in different parts of the brain. The neuronal hypothesis involved in this experiment is the general assertion that there is a significantly higher degree of *neuronal activity* in the hippocampal region when subjects respond to targets, i.e. stimuli they were shown before, in comparison to when foils are shown. What about the theoretical hypothesis? This is a hypothesis about how a cognitive process, such as memory or learning, works. In the study described here, the researchers seek to find out if the hippocampus is the brain region that realizes the neural processes necessary for performing the kind of recognition memory studied here. More generally, the theoretical hypothesis in this experiment is the assertion that the hippocampus is the neural substrate of recognition memory and retrieval.

Which of these hypotheses could be tested severely in this experiment? We can say that the hemodynamic hypothesis, namely that when subjects respond to targets

(stimuli they were shown before), there is an increased amount of hemodynamic activation in the hippocampal region, is the hypothesis we can severely test with this experiment. That is, we can check whether or not the inference $\mu_1 - \mu_0 > 0$ passes severely with the data we obtained from this experiment by looking at the error probabilities associated with the experiment. If it is the case that the experiment would probably yield a result that does not agree with this hypothesis, if it were false, then the inference $\mu_1 - \mu_0 > 0$ passes severely with this experiment. If so, then we have strong support for the hemodynamic hypothesis of interest. The crucial thing is that just the data that agreed with the hemodynamic hypothesis would not be enough for a severe test. An fMRI experiment yields a large data set, which is processed by a series of computational procedures and then analyzed using complicated statistical models and tests. The statistical tests come with a series of assumptions about the mechanism that generated the data and these assumptions constitute the statistical model that is used in the significance tests. It is necessary to check if these assumptions hold about the data generating mechanism in order to draw inferences reliably. It is also necessary to check for other potential errors, because in fMRI experiments many things at several different stages may introduce errors and these errors may create methodological flaws. In order to find out if we have reliable data it is necessary to check for errors such as biases in experimental design or biases in statistical analyses. If we have reliable fMRI data, we can then use those data to draw inferences about the hemodynamic hypothesis. For example, it is possible that the fMRI scanner used in this experiment was overly sensitive and this influenced the measurements it made. This would make the detection of a significant difference between control and experimental groups highly probable even if H_0 is true. If

so, we may be led to erroneously conclude that there is a genuine effect, while in reality this may just be due to the oversensitive measurements. This would make for a test of low severity for H_1: $\mu_1 - \mu_0 > 0$ that is, the value of SEV(T, $\mathbf{x}_0$, H_1) would be low.

In fMRI experiments, we need to check for such errors before we conclude that it constitutes a severe test of the inference of the form $\mu_1 - \mu_0 > 0$. We will address questions about experimental assumptions, data generating procedures, and statistical analyses in the next chapter when I closely study the experimental and data models of fMRI experiments. The most important point to note here is that for hemodynamic hypotheses we *can* check with what probability the fMRI experiment may lead us to an erroneous inference. This is not necessarily true for the neuronal or the theoretical hypotheses that are involved in fMRI studies, a result that I discuss further in the next section.

3.4: The Theoretical Significance of fMRI Findings

Thus far in this chapter, I have provided an account of the application of the hierarchical framework of models to fMRI experiments. I also showed how this account can help with underdetermination problems by clarifying the evidential import of fMRI results in a way that eschews overskepticism. We have seen that the "inferential distance" between fMRI results and substantive hypotheses can be covered by making the connection between the hemodynamic hypothesis of interest and fMRI results by framing the hypothesis in terms of statistical parameters, namely H_1: $\mu_1 - \mu_0 > 0$ when embedded in a statistical model of fMRI data. H_1 can then be tested against fMRI data by calculating the test statistic

$D = \bar{X}_1 - \bar{X}_0 / \sigma_{diff.}$. Indeed, we have already safely trekked from large-scale theories of human cognition to primary hypotheses that can be put to severe error probes in fMRI experiments. The next step is to go the remaining distance through experimental design and instruments of measurement to statistical analyses of data and neuroimages, which we can cover by continuing on the path of the models of inquiry with the notion of severe tests and error probabilities guiding the way. But, before that, let us note a few things regarding the other members of the primary models of fMRI studies, that is, the neuronal and theoretical hypotheses.

Well-received contemporary theories of cognition assume that cognitive processing is realized by neuronal activity, such as axonal action potentials or dendritic field potentials, while hemodynamic activity as such is not thought to be related directly to cognitive processing. The neuronal hypothesis in the Stark and Squire (2000) experiment was that there is going to be a significantly higher degree of *neuronal activity* in the hippocampal region when targets are shown compared to when foils are shown. This hypothesis cannot be severely tested by the fMRI experiment, because even though a significantly higher degree of hemodynamic activation was observed in the hippocampal region, this may have been due to factors other than high neuronal activity in that area, as stated by Huettel et al. (2004) and Logothetis (2008), and the fMRI experiment itself cannot test for this possibility. When applied to functional neuroimaging, the error-statistical approach avoids one related problem that was raised by Uttal (2001); namely that fMRI cannot distinguish between different cognitive processes that occur in the same brain region but realized by different spatio-temporal firing patterns of neurons. Since the hemodynamic hypothesis states nothing about

neuronal activity, it states nothing about spatio-temporal firing patterns of neurons. The hemodynamic activity observed in different trials of an fMRI experiment may have been caused by the same firing pattern of neurons or it may have been caused by a different pattern. In either case, the hemodynamic hypothesis still stands, because it has no commitments about any particular pattern of neuronal activity. So, we can talk about a robust effect of increased hemodynamic activation without committing ourselves to any specific, hypothesized pattern of firing of neurons. The robust effect of increased hemodynamic activation would be directly related to statistical inferences of the form $\mu_1 - \mu_0 > 0$ as discussed earlier. On the other hand, while it may be true that if there is a high degree of neuronal activity in a brain region then there is going to be a high degree of hemodynamic activity in that same region, the converse of this statement is not necessarily true. That is, if there is a high degree of hemodynamic activity in a brain region, this does not necessarily mean that there is a higher degree of neuronal activity in that region. At best, fMRI data can be used to falsify neuronal hypotheses but only in cases where no significant difference in hemodynamic activation is found between control and experimental conditions. However, the very reason that fMRI is used in cognitive neuroscience is to be able to test thoroughly hypotheses about brain function and falsifications alone do not contribute much to this endeavor. By focusing on what the fMRI data can reliably discriminate between, namely hemodynamic hypotheses, the error-statistical approach clarifies the evidential import of fMRI data. In addition, we can eschew overskepticism about fMRI, which stems mostly from such shortcomings, as its inability to discriminate between patterns of neuronal activity. By clarifying the

evidential import of fMRI data we can see clearly what it can contribute to cognitive neuroscience instead of dismissing it as a significant source of information.

The paradigm theory of cognitive science, or in Lakatosian terms the hard core of cognitive science, assumes a certain degree of modularity of cognition according to which human cognition is realized by distinct modules, each with its specialized cognitive function. Because of this general fact about cognitive science, the application of the hierarchical framework of models to fMRI has an interesting result apart from clarifying the evidential import of fMRI data. This can be seen when we think about the theoretical hypothesis in the Stark and Squire experiment, which states that the hippocampus is the neural substrate for recognition memory and retrieval. This statement certainly implies that it is assumed that there is a localized compartment in the brain whose function is required to realize recognition memory and that compartment is the hippocampus. Then, it is fairly easy to see that one important difference between the hemodynamic and theoretical hypotheses is that the theoretical hypothesis is stated in terms of theoretical concepts from cognitive psychology; namely recognition memory and retrieval. In using these terms, the theoretical hypothesis proposes the existence of certain cognitive functions that are separate from other cognitive functions and are localized in different parts of the brain, whereas the hemodynamic hypothesis is solely about hemodynamic activation as subjects perform cognitive tasks without assuming the truth of any large-scale, modularist theory of high level cognition. The fMRI experiment does not test for the existence of cognitive modules or functions as defined by theories of cognitive science. Also, as Huettel et al. (2004) suggest, researchers often disagree on what terms like 'retrieval' mean. When we conclude on the basis of an fMRI experiment

that retrieval, as it is defined by a modularist theory, is localized in a certain brain region, we may do so in error. This is because a separate cognitive function of retrieval as hypothesized by the researcher might not really exist, and yet the fMRI experiment cannot test for this error. The hemodynamic finding may be due to processes that are not necessarily those of a localized module whose function is retrieval.

The problem with the theoretical hypothesis discussed here is the same problem that has been raised by Uttal (2001) and by Hardcastle and Stewart (2002). The common thread in both works is the emphasis on the inherent circularity of assuming the existence of localized and well-defined cognitive modules prior to doing the fMRI experiment and then taking the results of the experiment as support for the modularist conclusions. Klein (2010b) recently voiced parallel methodological concerns for the low evidential value of fMRI data for what he calls functional hypotheses, which would correspond to the theoretical hypotheses discussed here. My account of the primary models in fMRI shows that when Uttal (2001) and Hardcastle and Stewart (2002) raise the circularity problem and when Klein (2010b) calls into question the evidential value of fMRI data for functional hypotheses, they are all talking about the theoretical hypothesis. In error-statistical terms, their criticisms end up saying the same thing, which is that the fMRI experiment is a test of low severity for the theoretical hypothesis, because it cannot check for the ways in which this kind of hypothesis can be ruled out. But here is an interesting question: Do our theoretical hypotheses have to assume any modularist theory of cognition? I do not think they do, because we can rephrase our theoretical hypotheses in more neutral terms. Instead of stating that "the hippocampus is the neural substrate for the cognitive function of recognition memory" which has a strong modular meaning; we

can state that "the hippocampus is involved in the recognition memory task used in this experiment." This rephrasing of the theoretical hypothesis at least to some extent eschews the strong modularity assumption. After all, to do an experiment on recognition memory does not necessarily commit one to the notion that there exists in the brain a module for recognition memory, or any other kind of memory for that matter. Moreover, as an experimental task, recognition memory is already a real process, because there is a clear operational definition for it which makes it possible to gather behavioral data from subjects. The term 'recognition memory' need not say anything more than what the subjects do as they perform the experimental task at least in the context of this experiment. In addition, the hippocampus is obviously a real anatomical entity in the brain. Thus, we can certainly use the neutrally restated theoretical hypothesis and test it in an fMRI experiment. The fMRI results can tell us something; at the very least it can give us a relationship between performance of recognition memory defined as the experimental cognitive task and a certain pattern of activation in the brain measured as the BOLD response. Furthermore, should we obtain this relationship, there would be concrete evidence in terms of our well-established theories of magnetic resonance and hemodynamics for the reality of this effect as was shown in chapter two. In the Stark and Squire experiment, there indeed is a relationship between activation in the hippocampus and performance of the recognition memory task. We can think of this kind of modest theoretical hypothesis as an unproblematic interpretation of the supported hemodynamic hypothesis, which is not in any way theoretically useless. In this way, we can get much more out of fMRI results than if we were to think of them as mere "first-pass sanity checks on experimental data" which can never "confirm functional hypotheses" as Klein

suggests we do (2010b; p. 275). Indeed, this is related to what Bechtel had in mind for the use of fMRI data in general. Bechtel suggests (2002) that fMRI data can provide constraints on theories of cognitive science. In order to provide constraints on theories, fMRI findings must be reliably established, but Bechtel does not go into detail about how fMRI findings are established or what they really show. The error-statistical approach offers a way of identifying what we take as the fMRI finding, specifically whatever the hemodynamic hypothesis states in an experiment, provided that the related inference $\mu 1 - \mu 0 > 0$ passes severely with the data obtained in that experiment. In the example discussed here, the established fMRI finding would be precisely what the hemodynamic hypothesis stated; specifically the general assertion that there is a significantly higher degree of hemodynamic activation in the hippocampus when subjects respond to stimuli they were shown before in comparison to novel stimuli. This finding can support the theoretical interpretation that the hippocampus is involved in recognition memory. In Bechtel's terms, any cognitive theory that excludes the involvement of the hippocampus in recognition memory would have to be revised or rejected.

In this chapter, I addressed the question 'what goes into the primary models of a typical fMRI experiment?' We have answered that question; the hemodynamic hypothesis is the kind of hypothesis that should be located in the primary model. When embedded in a statistical model of the data, the hemodynamic hypothesis of interest is framed in terms of the statistical parameters as the alternative hypothesis $H_1: \mu_1 - \mu_0 > 0$ which can then be placed in the severity function $SEV(T, \mathbf{x}_0, H_1)$ and severity assessments can be carried out to see whether or not we have genuine support for the hemodynamic hypothesis when we infer $H_1: \mu_1 - \mu_0 > 0$. The next question to ask is

whether, and how, we can put hemodynamic hypotheses to severe tests in fMRI experiments. This can also be thought of as an exercise in identifying the other arguments of the severity function, specifically, the test *T* and data $\mathbf{x}_0$. Of course, since any fMRI experiment involves a number of complicated inferential steps and procedures and we can think of test *T* as the whole procedure that generated the data, *T* would include several components. Thus, we need to see how we can break down an fMRI experiment into its components of experimental analogs and data processing and analysis. To do this, we have to identify the kinds of things that should be located in the experimental and data models of a typical fMRI experiment. As we do this, we can address in a piecemeal fashion the local issues and problems that may arise at different stages of the fMRI experiment, which may jeopardize the reliability of the inferences drawn from fMRI data. It is to this project that we turn in chapter four.

Chapter Four:

Experimental and Data Models in fMRI: Tackling Duhemian Problems

> One marked characteristic of the literature dealing with the cerebral circulation [of blood] is, we think, the contradictory nature of the results which have been obtained by different investigators. ... the cause of these discrepancies is to be found in the great difficulty of avoiding the sources of error which plentifully surround the subject, and in overcoming certain technical difficulties ... The ease with which one can obtain results upon certain points, on taking up the subject, is itself, we believe, apt to make the inquirer careless in controlling sources of error, which, it may be noted, are some of them not at first sight obvious.
>
> Roy & Sherrington, 1890 (p.85)

4.1: Introduction

Roy and Sherrington were the physiologists who postulated in 1890 the hypothesis that the volume of blood flow in the brain varies locally in parallel with changes in local functional activity. This hypothesis is one of the founding principles of cognitive neuroscience. As we discussed in chapter two, it took about a century of scientific developments and experimental progress in physics and physiology, which eventually made possible the introduction of fNI tools such as PET and fMRI, before cognitive neuroscience came to be a discipline. It is interesting to note that back in 1890 Roy and Sherrington anticipated some of the major problems that functional neuroimaging would face today. In the above quote, they point to problems such as the easiness of obtaining certain findings, discrepant results coming from different labs, and the difficulty of controlling for these errors. In chapters one and three, I have discussed authors who

emphasized these problems and consequently cast doubt on fMRI as a research paradigm. For example, Uttal (2001), Klein (2010a), and Roskies (2010) all point to some or all of those problems about which Roy and Sherrington have warned us. One major theme from Roy and Sherrington (1890) to Roskies (2010) is the challenging problem of identifying the sources of error in order for them to be controlled in various ways. Each component procedure of an fMRI experiment may introduce biases or errors and hence need to be scrutinized with respect to the kinds of errors they may introduce. In this chapter, I take on Roy and Sherrington's challenge of identifying some of these sources of error and propose ways in which they can be controlled.

In chapters one and three, I discussed generally how Mayo's notion of severe tests and the hierarchical framework of models of inquiry can be applied to fMRI studies. In chapter three, I addressed the question of what kind of hypotheses can be placed in the primary models of a typical fMRI study about which we can draw reliable inferences from data. This was an exercise in dealing with the general problem of underdetermination of hypotheses by data as it manifests itself in functional neuroimaging. In the end of this exercise, I concluded that hemodynamic hypotheses, specifically speaking, hypotheses about the amount or pattern of hemodynamic activation in different regions of the brain, should be the kind of hypotheses placed in the primary models. This meant that hemodynamic real effect hypotheses can be embedded in statistical models of data and framed in terms statistical parameters, e.g. $H_1{:}\mu_1 - \mu_0 > 0$. These hypotheses *can* then be put to severe tests against fMRI data.

Now that we know what kinds of hypotheses can be tested against fMRI data, the purpose of this chapter is to elucidate *how* this can be done. I propose a general

framework in which we can check, on a case by case basis, whether or not fMRI experiments constitute severe error probes of inferences to specific hemodynamic real effect hypotheses that are framed in terms of statistical parameters; as in H_1:$\mu_1 - \mu_0 > 0$. My strategy in constructing this general framework is to continue on formulating general questions in terms of the hierarchical framework of models of inquiry. The central question of the previous chapter was 'what should/can we place in the primary models of an fMRI study?' In parallel with this, the general question I ask now is this: 'what should/must we place in the experimental models and data models of a typical fMRI study?' Any fMRI study is a highly complex whole that yields a huge and noisy data set, and as such, a lot of procedures are required from preprocessing of raw data to statistical modeling and analysis. Naturally, things may go wrong during several of these procedures and this in turn may introduce biases or errors influencing the outcome of the experiment.

It is important to note here that the task of this chapter may be thought of as an exercise in dealing with Duhemian problems as they manifest themselves in functional neuroimaging. Duhem's problem arises when a scientist does an experiment, or a series of experiments, to test some hypothesis *H* and gets a result that does not agree with *H*. One construal of the problem is to think of it in terms of a *modus tollens* of the type Popper discussed: If hypothesis *H* is true, then the experiment will yield data *e*. Experiment yield not-*e*. Therefore, not-*H*. Of course, as an experimentalist philosopher of science using the hierarchical framework of models of inquiry, I deny this kind of simple rejection of hypotheses. In actual scientific practice, things quite often do not work this way. It is rather like this: If H_1, H_2, H_3,…, H_n and A_1, A_2, A_3,…,A_m, then *e*. Not-*e*.

Therefore, not-H_1 or not-H_2 …, or not-A_1 or not-A_2… where H_1 through H_n and A_1 through A_m are auxiliary hypotheses and assumptions involved in the experiment that yielded not-*e*. The latter inference is the only one that deductively follows. Thus, it appears as though that we do not know if it is the hypothesis that we should blame for not-*e* and falsify *H*, or if it is any of the auxiliary hypotheses or assumptions that is responsible for obtaining not-*e* and we should hang on to *H*. Now, we can think of H_1 as the hemodynamic hypothesis we place in the primary models of an fMRI experiment. Thus, H_2, H_3,…, H_n and A_1, A_2, A_3,…,A_m can be thought of as the auxiliary hypotheses and assumptions about the several different aspects of fMRI, from preprocessing procedures to statistical modeling and testing. For example, H_2 may be that variability in subjects' brain anatomy is not large enough to bias results, or A_1 may be that the type of fMRI scanner used is sensitive enough to detect any real effects. We can place all these aspects of an fMRI study in their proper locations in the hierarchical framework of models and find out how they may influence the experimental outcomes independently of the truth or falsity of the hemodynamic hypothesis of interest. That is, we find out their error characteristics or the error probabilities associated with their use, which are necessary for severity assessments of given fMRI studies, on a case by case basis, as tests of specific hypotheses such as H_1:$\mu_1 - \mu_0 > 0$. The outcome of these severity assessments in turn will inform our decisions about whether or not we have support for the hemodynamic hypothesis at hand.

It is granted that this chapter will not provide a complete solution of Duhem's problem, because I do not have enough space here to carry out the kind of analysis described above for each and every aspect of fMRI studies. Neither practitioners nor

philosophers of functional neuroimaging know all the ways in which fMRI procedures may introduce errors. In fact, a sizable portion of the literature is on fMRI itself; where practitioners and statisticians constantly try finding ways of improving the effectiveness of the procedures used to collect, model, and analyze data. In this chapter, I illustrate the general approach described above by placing certain aspects of fMRI studies in experimental and data models, and formulate ways in which we can find out how they may introduce errors in experiments and how this knowledge in turn can help us with severity assessments. In any given fMRI study, if we analyze some of its component procedures and find out that they do not influence experimental outcomes in any significant way, then we can rule out those procedures of the experiment as possible sources of error. Thus, in this way, Duhemian problems are resolved to the extent that Duhemian objections cannot be cogently raised about those specific procedures of the fMRI experiment, which produce errors with low probability and so we can rule out as sources of error . We start with experimental models in the next section.

4.2: Experimental Models

If we recall from chapter three, the experimental models provide the link between the data collected in the experiment and the primary hypothesis being tested. The experimental model serves two functions: the first function is to provide "a kind of experimental analog of the salient features of the primary model" (Mayo, 1996; p.133). If the primary problem in an experiment is testing a hypothesis, then the experimental model tells us what is expected to obtain in this experiment with high probability if the hypothesis is true. As can be seen in figure 4.1, we have said that the main thing in the

experimental model is the statement of the predicted result in the scenario that the primary hypothesis is true. Hence, in a typical fMRI experiment, where the primary hypothesis is of the form 'brain region X is involved in the performance of cognitive task C', the experimental model tells us that as subjects in the fMRI scanner perform an example of the cognitive task C, the scanner will with high probability register a significantly high amount of activation in region X of their brains.

The second function of the experimental model is "to specify analytical techniques for linking experimental data to the questions of the experimental model" (ibid., p.134). Because of the many sources of error that influence the data collection process, the data will very rarely agree exactly with the researchers' prediction. To deal with this, the experimental model may statistically formulate the link between the primary hypothesis and the data model. Thus, the first job of the experimental model is to link the primary model (primary hypotheses) with the experimental model (experimental hypotheses), and its second job is to link the experimental model (experimental hypotheses) with the data model (experimental data).

In this section, I will discuss two aspects of fMRI studies the discussion of which corresponds roughly to the two functions of the experimental models; the first has to do with the central component of any fMRI study, namely the fMRI scanner and its characteristics. Recall that the predicted outcome of any fMRI study is stated in terms of what the fMRI scanner will register as observed hemodynamic activity as subjects perform cognitive tasks, so scrutiny of the characteristics of the specific type of scanner used in an fMRI study is of crucial importance for any assessments of errors that may be introduced by the scanner. The second aspect has to do with anatomical variability across

subjects in an experiment. People's brains vary with respect to size, shape, and orientation of anatomical landmarks, etc. Because of this, certain regions of the brain may be misidentified in some subjects, which will lead to errors in establishing causal relationships between cognitive activity and brain regions. Thus, researchers need to assess how probable it is to commit such errors, which can be done by providing statistical measures of variability in brain anatomy across subjects. Both of these aspects of fMRI are examples of how experimental models provide the links between the primary model and data models of fMRI studies. Before moving on, one thing has to be noted. Throughout this chapter, there is an ambiguity in the use of the term 'brain region'. This term may refer to an anatomically identified structure, as in talking about the hippocampus which has anatomically distinct boundaries that make it fairly easy to identify. Or, the term brain region may refer to a group of voxels identified as one region because they are all found to be activated as subjects perform a specific cognitive task. This is more common when we talk about the cerebral cortex where anatomical boundaries are not as clear as they are in the limbic system. This ambiguity is a characteristic of the field of cognitive neuroscience. One reason is that for obvious reasons we cannot dissect and stain human brains, so we have to rely on fMRI results to identify regions functionally. The other reason is that cognitive neuroscience is still too new a field. We simply do not have the knowledge to produce a clear-cut functional compartmentalization of the human brain. Indeed, this is the overall aim of cognitive neuroscience. To get there, researchers make use of every bit of information they can obtain, sometimes defining brain regions anatomically and sometimes functionally.

4.2.1: Power of fMRI Scanners As A Source of Error: The fMRI scanner works by generating a powerful magnetic field. In an fMRI experiment, as subjects perform cognitive tasks, those regions in the brain that are involved in doing the cognitive tasks need more energy, which is supplied by oxygenated hemoglobin. In active regions of the brain, there results an imbalance between oxygenated and deoxygenated hemoglobin. Since oxygenated hemoglobin and deoxygenated hemoglobin have different magnetic properties, this imbalance leads to inhomogeneities in the magnetic field around the activated regions of the brain. By applying a strong magnetic field to the chamber inside the scanner, fMRI detects these inhomogeneities in the magnetic field as the blood-oxygen-level-dependent (BOLD) response. There are two fundamental problems in this process: the first problem is that the measured change in the BOLD signal when subjects perform a task is very small when compared to the total intensity of the magnetic resonance signal (Huettel et al., 2004; p. 217). In other words, the change in the BOLD signal between control conditions, i.e. when subjects are in rest, and experimental conditions, i.e. when they perform a task, is very small. As a result, what the scanner detects as the difference between conditions of an experiment, that is, the fMRI finding, is an absolute but very small effect. The ratio of the intensity of the task-related fMRI signal and the general variability of the signal due to all sources of noise yields a very small value.

The second problem has to do with variability in the BOLD signal over time. Several factors affect the variability in the BOLD signal. For example, the temperature of the subject's body while in the scanner, subject's head motion, heart rate, and respiration are all factors, other than the task being performed, that influence the variability of the

signal. The change in the BOLD signal that is related to the task is very small when compared to the total variability of the signal. Consequently, in any fMRI experiment, there is the danger of the task-related change in the BOLD signal being masked by other sources of variability, or in other words, the signal of interest may be lost in the noise (ibid.). Because of this, some fMRI experiments may end up lacking sufficient power to detect signals of interest. Practitioners of fMRI deal with this problem by defining the functional signal-to-noise ratio (SNR). Here, the signal is defined as the difference in fMRI data between two states of brain activity hypothesized to be caused by an experimental manipulation, for example, between control (no task) and experimental (task) conditions. Noise is defined as the variability in fMRI data in these states of brain activity over time. Then, the functional SNR is the ratio between these two quantities (ibid., pp. 220-221). The higher the functional SNR, the easier it is to detect task-related changes in fMRI data.

Huettel and his colleagues (2004) talk about several different ways of improving the functional SNR, one of which is to use fMRI scanners that generate stronger magnetic fields. The strength of a magnetic field is measured in terms of the *Tesla* (T) unit; the strength of earth's magnetic field is 0.00005T. In cognitive neuroscience, fMRI scanners with magnetic field strengths from 1.5T to 7.0T can be employed, so the magnetic field an fMRI scanner generates is very strong (Lazar, 2008; Hashemi et al., 2010). As Huettel et al. (2004) state, one primary factor determining functional SNR is net magnetization. Net magnetization is proportional to the strength of the magnetic field; the raw SNR increases roughly linearly with field strength. Therefore, as researchers use scanners of higher magnetic field strength, the functional SNR is improved.

However, one issue with increased magnetic field strength is that at higher field strengths more noise is detected by the fMRI scanner in addition to task-related BOLD signal. For example, physiological noise, which is the variability in the BOLD signal due to heart rate and respiration, increases quadratically with field strength (Huettel et al., p.239). Consequently, as field strength is increased there is greater danger of the fMRI scanner detecting noise as if it is a real effect. Another danger is that when scanners of higher field strengths are used, task-related changes in the BOLD signal may be lost in increased physiological noise, as Huettel et al. write; "as field strength increases above about 4.0T … increases in physiological noise may counteract gains in [task-related] signal, setting an asymptotic upper limit for functional SNR" (ibid.). Thus, at field strengths higher than 4.0T we may get too many cases where noise is detected as real effect. Indeed, this is noted by Savoy (2001) as a concern about fMRI studies in general. Savoy argues that this is very similar to the problem that Meehl (1967) raises, which was discussed in chapter one. Meehl's claim was that if we sufficiently increase the power of statistical significance tests, then we can reject any null hypothesis even if the null hypothesis is exactly true. Likewise, it appears that if we increase the field strength of our fMRI scanner, we may find evidence for hypothesized relationships between certain cognitive processes and certain patterns of brain activation regardless of whether or not those relationships do in fact hold. In other words, the observed activation may simply be noise and may have nothing to do with the performance of a cognitive task. At higher field strengths we may end up having oversensitive measurements where we detect noise due to overall activity of the brain as real effects. If this is true, then such fMRI experiments would be tests of low severity of hemodynamic real effect hypotheses and

inferences to H_1: $\mu_1 - \mu_0 > 0$ where μ_1 is mean hemodynamic activation in the experimental condition,when subjects perform cognitive tasks, and μ_0 is mean hemodynamic activation in the control condition, when subjects are at rest. This is because with high probability they would yield data that agree with the alternative hypothesis even if there was only a trivial discrepancy from the null hypothesis. Thus, in general, one could argue that increasing field strength increases probability of errors and, as a result, experiments using high field strength scanners do not constitute severe tests of inferences like $\mu_1 - \mu_0 > 0$. We can use such severity assessments as a guide in evaluating the possibility that we have a real effect as opposed to an artifact of the fMRI scanner.

Let us assume that we do two fMRI experiments, where the only difference between the experiments is the fMRI scanner used, say in one experiment we use a 3.0T scanner and in the other experiment a 7.0T scanner. We test the same hypothesis in both experiments, which states that there is going to be a high amount of activation in brain region X when subjects perform cognitive task C. This hemodynamic real effect hypothesis can be embedded in a statistical model and framed in terms of statistical parameters: H_1: $\mu_1 - \mu_0 > 0$. Everything else in the experiments is also the same but the scanner used. If in both experiments we obtain significantly high amount of activation in X when subjects perform C, then one could argue that the experiment that used the 3.0T scanner was a more severe test of H_1: $\mu_1 - \mu_0 > 0$ than the one that used the 7.0T scanner. This is because the chances of detecting noise as a real effect are greater for the 7.0T scanner than they are for the 3.0T scanner. Thus, it may be argued informally that if the inference being tested is false, the probability of the experiment with the 3.0T scanner yielding data that do not agree with the hypothesis is higher than the experiment with the

7.0T scanner. So, the results from the 3.0T scanner would be more indicative of a real effect than results from the 7.0T scanner. The severity assessment here comes from a consideration of characteristics of fMRI scanners of different field strengths based on scientific facts about magnetic field strength and the BOLD signal. This type of severity assessment is useful as one of the criteria in evaluating fMRI results with respect to whether or not they constitute real effects. Of course, having the proper error probabilities between 0 and 1 would be a much more effective way of controlling for this error. How can one do that?

In essence, this is not very different from calculating false positive and false negative rates for diagnostic tests, for example, as for tests for diagnosing patients for certain viruses or diseases. In medical cases, we may have more than one way of testing for the disease, so the results of several different types of tests may be used in calculating false positive and false negative rates of another test. In the case of fMRI, one may use results of animal experiments to assess false discovery rates of scanners of different field strength. For example, if we know from animal experiments that a certain brain region is involved in some learning task, then we can do fMRI experiments using scanners of different field strengths where subjects would do the same learning task in the experimental condition and in the control condition. Then, we can look at how often different types of scanners detect activity in the control condition where no effect is supposed to be obtained. However, many fMRI experiments are done with human subjects where higher cognitive functions, such as language or decision making, are of interest. Since we do not have animal models of language processing, we cannot obtain any knowledge about such functions from animal studies. Nonetheless, there is another

fairly straightforward way of assessing error probabilities of fMRI scanners. One could collect data sets from many different fMRI studies that have used scanners of different field strength. Then, a group of researchers and statisticians can be asked to do statistical tests on data sets that come only from the control groups of these studies without letting them know that the data come from control groups. Thus, all the data would be of subjects at rest. After the statistical tests are carried out, one can look at the proportion of the significant effects they report from these data sets. This proportion would give us something that is close to a false positive rate for each type of fMRI scanner. On the basis of the informal severity assessments above, one can predict that the false positive rate of 7.0T scanners would be higher than the false positive rates of 1.5T or 3.0T scanners. In fMRI experiments, either with informal assessments of error characteristics or actual error probabilities associated with scanners of different field strengths, researchers can evaluate the possibility of the error of detecting noise as real effects. If these probabilities are sufficiently low, then researchers can reasonably rule out errors of this kind.

In any given experiment, one of the assumptions is that the probability of the fMRI scanner detecting noise as a real effect is low. This assumption can be identified as one of the auxiliary assumptions in the above formulation of Duhem's problem. Thus, when we scrutinize the scanner used in this experiment with respect to its field strength and if it is a scanner of 3.0T, then we can say that the probability of error due to field strength is low and rule this out as a source of error. Of course, there may be factors other than field strength that may lead to detection of noise as real effects, such as multiple testing or thresholding problems. Since these issues come up in the modeling and analysis of data, I will discuss these problems in section 4.3 which is on data models. Let us now

look at another feature of fMRI studies that can be located in experimental models of the hierarchical framework of models.

4.2.2: Dealing With Neuroanatomical Variability: The essential type of inference in fMRI studies is about where in the brain, if anywhere, there is significantly increased activation. This kind of inference is mostly drawn across subjects; it can take the form "subjects in the experiment had significant activation in brain region X." Let us remember from chapter three that the hemodynamic hypothesis is what is tested in fMRI experiments. This hypothesis is usually of the form 'when subjects perform cognitive task C, there will be a significantly higher amount of activation in brain region X than when they do nothing.' Thus, in any fMRI experiment, if we can reliably infer from data that subjects in the experiment had significant activation in brain region X, and not any other brain region, when they performed C, then we have evidence for the truth of the hemodynamic hypothesis. At first glance, this kind of inference seems to be unproblematic assuming that other aspects of the experiment, such as statistical modeling and analyses, are free of biases or flaws. However, there are some issues that need to be addressed. Huettel et al. (2004) raise some of these issues: 'how do neural activity map onto neuroanatomy?' 'How consistent is that mapping across subjects?' 'How do functional data "correspond" to underlying neuroanatomy?' To address these questions, fMRI data have to be mapped onto high resolution structural images. Here, we have to remember the fact that people's brains vary with respect to size, shape, orientation, and gyral anatomy. Brain sizes of two subjects in a given fMRI experiment may differ by 30 per cent. A hidden assumption in fMRI data analyses is that each voxel (a volumetric

pixel corresponding to a very small chunk of brain tissue) represents a unique and unchanging location in the brain and this assumption is always wrong (ibid., p. 253). For example, voxel M may correspond to region X in one subject while the same voxel may correspond to region Y in another subject. Also, brain shapes of people differ a great deal, for example as in long and thin versus short and fat brains. The organization of sulci and gyri is also variable across individuals in ways that major landmarks in the brain may be at different positions and may be differently oriented across individuals. Therefore, when we draw an inference of the form "subjects in the experiment had significant activation in brain region X" we do not know, because of neuroanatomical differences, whether or not in each subject the activation is really in region X. For a given subject B, it may be true that there is significant activation somewhere in her brain, but it may be in region Y, which in her brain is just adjacent to region X. That is, what corresponds to region X according to the mapping used by the fMRI procedure may in fact be region Y in subject B's brain. This is problematic for any generalizations that associate a brain region with the performance of some cognitive task. For example, let us say that in an experiment subjects were asked to perform cognitive task C and the results show that they had significant activation in brain region X. We conclude that region X is involved in the performance of cognitive task C. However, subject B, whose brain anatomy differs from other subjects, performed the same cognitive task but the activation may have been in region Y of her brain. While we may assume that she had significant activation in region X, because we did not take neuroanatomical variability into account, we do not really know if it was region X or region Y that was activated. Moreover, anatomical variability in other subjects' brains may also complicate our inferences. Therefore, our

generalization may be in error and we do not know if we committed this error or how probable it was that we committed it in this experiment.

To address this problem, researchers apply a procedure called normalization in which shape differences across brains are compensated for by mathematically stretching, squeezing, and warping each brain so that it is the same as other brains. In most normalization procedures the Talairach stereotaxic space is used, which is a coordinate system of the brain that defines locations of brain structures in terms of their coordinates (Talairach & Tournoux, 1988). The brain that was used by Talairach and Tournoux to develop this system was that of an elderly lady, which creates problems of representativeness. Subjects in fMRI experiments would probably have brains that are different from the brain that is taken as a model by the Talairach space. The probability of drawing false inferences may be reduced by normalization. However, since we do not have empirical measures of the variability across brains and the representativeness of the brain used in the Talairach space is questionable, we cannot safely assume that errors due to neuroanatomical variability are sufficiently reduced. In other words, we do not have the accurate, or even approximate, error probabilities associated with inferences of the form "subjects in the experiment had significant activation in brain region X." But, just like the error probabilities of fMRI scanners of different field strengths, we need the error probabilities stemming from neuroanatomical variability to be able to control for this kind of error. This is one of the steps in the overall severity assessments of fMRI experiments as tests of hemodynamic real effect hypotheses.

In order to control for and minimize errors of this kind, some scientists suggest using probabilistic spaces based upon combining data from hundreds of neuroanatomical

scans. One probabilistic space used in normalization is the Montreal Neurological Institute (MNI) template based on hundreds of brain images (Mazziotta et al., 1995). This a step toward approximating more closely the actual error probabilities associated with inferences to hypotheses about relationships between cognitive performance and activation in certain brain regions. Of course, there may be biases in this atlas and there is always room for improvement; in fact other groups of researchers are currently working in collaboration with the Mazziotta group for updates. Duncan (2009) in the Discover magazine reported that the team of researchers studied scans of 450 brains and used hundreds of thousands of images taken from 7,000 people around the world as they updated the atlas.

Let us assume that we do an fMRI experiment on the neural substrates of working memory. Previous studies have shown that the caudate nucleus, a brain structure that is connected with the thalamus and higher cortical structures, is involved in certain working memory tasks (Baier et al., 2010; Provost et al., 2010). We test the hypothesis "the caudate nucleus is involved in the performance of working memory task W." The results of our fMRI experiment show that when subjects perform W, there is significantly high hemodynamic activity in a group of voxels that we identify as the caudate nucleus. Therefore, assuming that the experiment was carried out without any serious flaws, we conclude that the caudate nucleus is in fact involved in the performance of working memory task W. One could object to this conclusion and say that this result does not necessarily mean that there really was activation in the caudate nucleus. This is because it can be argued that some subjects' brains may have been sufficiently different anatomically that although the results show activation in the caudate nucleus as is

identified by the Talairach space, perhaps several subjects had activation in the internal capsule, a brain structure adjacent to the caudate nucleus. Such an objection can be addressed if we have used a probabilistic brain atlas. That is, as we analyze our data we can take into consideration the error stemming from anatomical variability. The probabilistic brain atlas would give us the probability of correctly identifying a group of voxels as a specific structure on the basis of hundreds of brain scans. In the case at hand, we are interested in activity in the caudate nucleus, so we can consult the probabilistic atlas about the group of voxels we found to be activated. The atlas tells us how often that group of voxels has been identified as the caudate nucleus; specifically, 92% of the time the group of voxels has been identified as the caudate nucleus, 6% of the time as internal capsule, 1% as anterior horn of lateral ventricle, and less than 1% as other regions (Mazziotta et al., 1995). This information can help us find out how often we may commit the error of misidentifying of brain regions. We can say that the probability of committing the error of misidentifying the group of voxels in this experiment as the caudate nucleus was 8% and this probability has come from hundreds of scans. Very rarely fMRI experiments have more than 15 or 20 subjects, so the probability of misidentification may be even lower in our experiment. This is because, the more subjects we have the more anatomical variability we have that could produce errors in identification of brain regions. In the worst case, the probability of misidentifying the group of voxels was 8%, which is not too high. So, with the error probability associated with this type of error at hand, we can say that the chances of committing this error are low. Therefore, we can rule this out as a serious source of error and we have a more reliable way of inferring accurately where in the brain activation really took place.

Again, as with the field strength example above, we can think of one of the assumptions in the above formulation of Duhem's problem to be about neuroanatomical variability. In our experiment, this assumption would state that we do not misidentify brain structures of interest with high probability. We can look at the probability of misidentifying the caudate nucleus and if this probability is not too high, then we can reasonably say that this assumption was not a big source of error. In general, it is granted that neuroanatomical variability may introduce errors, but in a given experiment, if the probabilities associated with this kind of error are assessed, as is done in probabilistic brain atlases, and if these error probabilities are found to be low, we can say that it is improbable that this specific error was committed in the given experiment.

In this section, I discussed two different aspects of fMRI studies, namely the use of fMRI scanners of different magnetic field strengths to improve the functional signal-to-noise ratio (SNR), and anatomical variability of the brains of subjects in fMRI studies. My major focus was on the error characteristics associated with these aspects, how they may lead to errors, and how these sources of error may be addressed with the help of error-statistical notions of error probabilities and severe tests. In terms of the hierarchical framework of models of inquiry, I identified two aspects in the experimental models and analyzed how they may influence experimental outcomes and how related errors can be controlled (see figure 4.1).

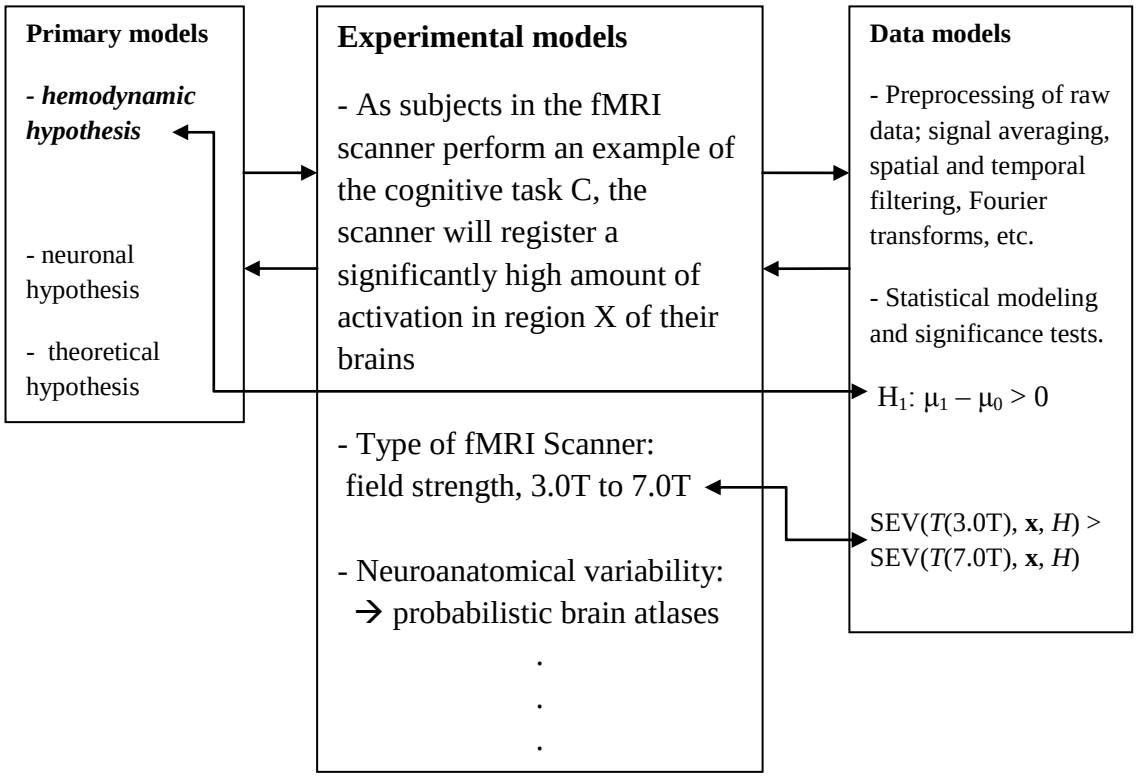

Figure 4.1: Experimental Models With Type of fMRI Scanner and Neuroanatomical Variability As Sources of Error

In both cases of using scanners of different field strengths and neuroanatomical variability, having the error probabilities associated with these aspects of fMRI would be of great help in carrying out severity assessments of fMRI experiments. One thing to note here is that in the case of magnetic field strength we had to do with informal scrutiny of error characteristics of different types of scanners, because we do not have adequate calculations of actual false discovery rates.

In the case of errors due to neuroanatomical variability, we have actual probabilities of misidentifying brain regions thanks to practitioners who saw this issue as a potential source of error and started projects of assessing anatomical variability across subjects and calculating probabilities of committing this kind of error. Indeed, this is a

common theme in contemporary functional neuroimaging; a sizable portion of the literature looks into types of errors in fMRI studies and tries to find ways of dealing with these errors. The development of probabilistic brain atlases is an example of this current trend and many other examples can be found in Lazar (2008). As studies like these are completed, we will have error probabilities associated with other component parts of fMRI studies, which can all be incorporated into severity assessments of experiments. In other words, more and more component parts of fMRI studies can be placed in the experimental models and their error characteristics can be analyzed. Statements about these component parts may be thought of as corresponding to auxiliary hypotheses and assumptions in the formulation of Duhem's problem in the beginning of this chapter. If we find out that the error characteristics or error probabilities associated with these component parts are low, then more assumptions can be ruled out as serious sources of error. The component parts of fMRI studies can also be thought of as aspects of the data generating mechanism, i.e. the test T in the severity function SEV(T, $\mathbf{x}_0$, H). As such, as we learn about the error probabilities of these component parts, we can have more complete evaluations of the epistemic value of fMRI findings in terms of severity assessments. Therefore, as experimental knowledge about the instruments and procedures of fMRI accumulates, we will have a better and more complete understanding of how we can learn from fMRI studies. However, for any understanding of how we can learn from fMRI we have to look at another series of aspects of fMRI studies, which can greatly influence experimental outcomes and/or introduce errors. These are aspects of the statistical modeling and analysis of fMRI data which can be placed in data models in the hierarchical framework of models of inquiry.

4.3: Data Models

The data models provide the answers to two types of questions: the 'before-trial' question and the 'after-trial' question. The before-trial question is 'how should data be collected and modeled so that we can put the data in canonical form in order to be linked to the experimental model?' The after-trial question is 'how can we check whether or not the procedures of data collection were in line with the assumptions of experimental models?' As Mayo (1996) writes "data models, not raw data, are linked to the questions about the primary theory, and a great deal of work is required to generate the raw data and massage them into the canonical form required to actually apply the analytic tools in the experimental model" (p. 135). Thus, in order to have a useful philosophical account of inference in functional neuroimaging, we have to have a good understanding of how the data models are generated. These models provide the data sets on which statistical tests are carried out to draw inferences about the hemodynamic real effect hypotheses of interest. In the case of an fMRI experiment, several procedures from the preprocessing of raw fMRI data to statistical modeling and analysis of data come into the picture, which need to be placed in data models and scrutinized with respect to their error characteristics. We start with preprocessing of raw fMRI data in the next section.

4.3.1: Preprocessing of fMRI Data: The first thing to note in discussions about statistical analysis procedures in functional neuroimaging is that even the simplest fMRI experiment yields an immensely complex and large data set. The goal in an fMRI experiment is to relate changes in hemodynamic activation over time to an experimental

manipulation. Data are collected as a time series, a large amount of data on the hemodynamic processes in the subject's brain are acquired in temporal order at a specified rate as the subject performs some cognitive task, such as looking at a picture, reading a word, etc. Each session consists of multiple runs of presentation of the cognitive task and each run includes single images of the brain called volumes. Volumes consist of images of slices of the brain and slices consist of three-dimensional elements called voxels. A matrix of voxels makes up the slice where the matrix may be of size 64x64 or 128x128. In an experiment that studies the entire brain there may be as many as 25 slices. For example, in an experiment where the size of the voxel matrix is 64x64 and there are 25 slices in the volume, there would be time series data from a total of 102,400 voxels to be processed and analyzed. The fMRI data set can be thought of as a four-dimensional matrix; voxels by voxels by slice by time. In a simple 6-minute run of an experiment that covers the entire brain and where the fMRI scanner delivers an excitation pulse every second, the four-dimensional matrix of data be 64x64x25x360, where 64x64 is the size of the voxel matrix, 25 is the number of slices, and 360 is the number of volumes since data from the entire brain are recorded every second (Huettel et al., 2004; pp.186-188). Because of the complexity and the large size of raw data sets, several computational procedures, collectively called “preprocessing,” are needed to obtain data sets that can be put to statistical significance tests. Huettel et al. (2004) state that these preprocessing steps have two major goals; first, removing non-task-related variability from data and in this way increasing the signal-to-noise ratio, and second, preparing the data for statistical analyses.

Earlier, I discussed two fundamental problems in fMRI as an experimental paradigm; one problem was that the measured change in the fMRI signal is very small when compared to the intensity of the MR signal as a whole. The other problem was that the signal change in response to the experimental manipulation, that is, the cognitive task performed by the subjects, is very small compared to the total variability across data. These problems create the difficulty of identifying the changes in activation that are due to the performance of a cognitive task. In addition, we have to identify which changes in the signal are due to sources of variability other than the experimental manipulation, such as physiological noise (heart rate, respiration, etc.). These non-task-related sources of variability lead to data sets with high variability and, consequently, the task-related change may be lost in the noise or the overall variability of the data. To prevent this from happening researchers try to increase the functional signal-to-noise ratio (SNR). One way to improve the SNR is to use fMRI scanners that generate stronger magnetic fields, such as going from a scanner of 1.5T to one that generates a field of 4.5T. This may lead to certain errors, but increasing field strength is only one of the ways in which the functional SNR can be improved. Functional SNR can be improved either by increasing the fMRI signal, which can be done by using a scanner of stronger magnetic field, or by reducing the noise, which can be done in various ways such as signal averaging or applying temporal or spatial filters to raw fMRI data (ibid.). In either way, the functional SNR is improved. Here, I discuss spatial filtering, also known as smoothing, as a preprocessing procedure in more detail.

Spatial filtering, or smoothing, is a computational procedure applied to raw fMRI data in order to reduce the noise due to non-task related sources of variability such as

heart rate or respiration. If successful, one effect of smoothing is that noise is averaged out while the task-related signal is left unaffected (Lazar, 2008; p.48). Essentially, smoothing combines and spreads the data observed in multiple voxels, which ends up "blurring" the neuroimages. The fMRI signal as measured across voxels exhibit spatial correlations, that is, if a voxel is active, then with high probability nearby voxels are also active. There are many things that may be the reason for this; one probable reason is that adjacent regions of the brain may also be functionally similar. In addition, brain regions are highly connected with nearby regions, so when one group of voxels are active, this causes nearby voxels to be also active. Thus, using a spatial filter, which corresponds to spatial correlation expected to occur because of functional similarity and connections of brain regions, greatly improves the functional SNR (Huettel et al., 2004; p.277). A common blurring technique is applying a Gaussian filter, which can be characterized by the measure called "full width at half maximum (FWHM)" of the observed signal, which is defined as $2\sqrt{2\log\sigma}$ for a Gaussian distribution that has variance σ^2 (Lazar, 2008; p.48). When a Gaussian filter is applied to fMRI data, it spreads the observed signal over other voxels that are nearby. Spatial filters may be wide or narrow; narrow filters combine data from a few voxels, whereas wide filters combine data across many voxels (Huettel et al., 2004; p.276). The width of the spatial filter applied in an experiment is expressed in millimeters at half the maximum value of the fMRI signal. For example, a filter width of 10 mm FWHM combines data from approximately 3 voxels (Lazar, 2008; p.48). As the width of the spatial filter increases, more smoothing is applied combining data from more voxels. Spatial filters of 8 mm are commonly used in fMRI studies, but the spatial filter chosen depends on how good the SNR is without smoothing. That is, if

the SNR is bad, a wider filter is needed, whereas if the SNR is good enough, a filter narrower than 8 mm may work well.

There are two benefits of applying spatial filters (Huettel et al., 2004; Lazar, 2008). One benefit is that spatial filtering improves the functional SNR, thus making the fMRI experiment more powerful in detecting task-related signals. As Huettel et al. (2004) and Lazar (2008) note, the other benefit of spatial filtering is that it makes the data have a distribution closer to a normal distribution. Thus, spatial filtering is supposed to improve the quality of data for statistical analyses. In a volume of fMRI data there may be as many as 102,400 voxels. If the threshold for significance is set at .05, then, when we carry out significance tests for each voxel to determine whether or not it is active, assuming independence of voxels, as many as 5,000 voxels may be detected as active due to mere chance. If spatial filtering is applied, a smaller number of groups of voxels are found to be above the threshold for significance. Thus, as Huettel et al. (2004) argue, spatial filtering is helpful also for the multiple testing problem.

As with any other procedure employed in fMRI studies, spatial filtering has certain disadvantages. As Lazar (2008) notes, researchers have to be very careful in choosing the width of the spatial filter they will employ. There are several ways in which spatial filtering may lead to errors. If the width is not appropriate, the filter applied may have negative effects on statistical analyses of preprocessed data. If the filter employed is too wide, that is, data from many voxels are combined, then data from regions that are not active may be included. This may occur when there is significant activation in a very small brain region, but if data from nearby nonactive voxels are combined in the filter, then the activation in the small region may be smoothed out and rendered undetectable.

There may be cases where the activation in a small brain region is related to the cognitive task performed by subjects and, if too wide a spatial filter is applied, this functionally significant activation may go undetected. On the other hand, if the width of the filter is too narrow, it will not be effective in improving the SNR. Thus, nothing would be gained while spatial resolution would be decreased because any spatial filtering degrades spatial resolution to some extent. So, a filter that's too narrow may do only harm. Another disadvantage of spatial filtering is that it may cause the merging together of brain regions that are functionally different. That is, regions that have been shown to be involved in different types of cognitive tasks in previous neurobiological experiments. This kind of merging may lead to contradictory fMRI findings in different experiments or even in different kinds of analyses of the same data set. All these disadvantages of applying spatial filters to fMRI data may introduce errors that affect the outcomes of statistical analyses of data. Thus, they may influence the experimental findings independently of the truth or falsity of the hypothesis that the fMRI experiment is meant to test. It is possible for fMRI researchers to obtain results that agree with their hypothesis not because the hypothesis is true, but because they applied a spatial filter to their data.

Fransson et al. (2002) demonstrated how this can happen. In an fMRI experiment Fransson and his colleagues asked the subjects to do an episodic memory encoding task as fMRI data were collected. They then applied two different types of analyses to the same data set; one type of analysis included no spatial filtering, and the other included a spatial filter with a width of 8 mm, a filter size commonly used in the fMRI literature. The analysis with spatial filtering yielded significantly high amounts of activation in the hippocampus, whereas the analysis without spatial filtering did not yield high activation

in the hippocampus. This result demonstrates how an fMRI finding may be obtained because of some procedure applied to data in the preprocessing stage rather than the truth of the hypothesis being tested. As Franssen and his colleagues state, "the results of an fMRI study appear to be crucially dependent on the approach chosen for post-acquisition data processing and analysis" (ibid., p. 981).

Because of these disadvantages of spatial filtering, Lazar (2008) notes that some fMRI research groups do not include any spatial filtering as a part of their preprocessing protocols. This may be too radical a choice; because, as was discussed above, not applying any spatial filters may lead to an experiment with low power where any task-related signals of interest may go undetected. As with any other aspect of the fMRI experiment, trying to find out the error characteristics of spatial filtering is a better approach than either applying spatial filters blindly or not applying any spatial filter at all. Lazar advocates an approach in a similar vein. She suggests analyzing fMRI data sets without spatial filtering and then analyzing the same data sets several times with spatial filtering of varying widths. The results of these analyses can show us how dependent the experimental results are on spatial filtering. This could also tell us how often the inferences we draw from data are influenced by procedures like spatial filtering. Indeed, analyses of this type must be expanded to include other preprocessing steps, such as temporal filtering, etc. Ideally, each preprocessing step would be analyzed with respect to its error characteristics and how it may influence the experimental outcomes. This approach has been taken by only a few groups of researchers.

One paradigm for the assessment of the effects of preprocessing procedures has been proposed by Strother and his colleagues. The paradigm is called the "nonparametric

prediction, activation, influence, and reproducibility resampling" or NPAIRS framework (Strother et al., 2002; LaConte et al., 2003). This paradigm makes use of the notion of cross-validation where an fMRI data set is split in two halves; one half is designated as the "training" data and is used to estimate the parameters for a predetermined model. The estimated parameters and the model are used to make predictions to be tested on the other half of data, which is designated as the "test" data. This process is repeated in a second run but with the training and test data switched, that is, in the second application of the process test data are used for training and training data are used for testing. In this way, researchers assess the prediction accuracy of their models. Reproducibility of the experimental findings is assessed by comparing the results of statistical analyses on both halves of the data across several runs. The flexible nature of this analysis paradigm allows researchers to assess the effects of different types of preprocessing protocols on fMRI data that are subjected to statistical analysis. For example, LaConte et al. (2003) compared the effects of different preprocessing protocols on prediction accuracy and reproducibility. Across several runs of the split half process described above, they applied different preprocessing protocols, which they called analysis chains. Each analysis chain includes different levels of preprocessing of the fMRI data. For example, one chain included no preprocessing procedures, whereas others included normalization and different degrees of spatial filtering, for example, one chain applied a narrow filter and another chain applied a wide filter. Then, they did final statistical analyses on data sets that came from these different analysis chains in order to assess the effects and contribution of different preprocessing protocols, or analysis chains, on prediction

accuracy and reproducibility. The results showed that spatial filtering (smoothing) was the most effective procedure in improving prediction accuracy and reproducibility.

However, as LaConte and his colleagues note, there are no general pre-data guidelines for what the optimal preprocessing protocol would be for all experiments (2003). One reason for this is that the optimality of a preprocessing protocol is dependent not only on the elements of the protocol, as in how much smoothing or normalization was applied, but also on other experimental parameters such as the type of scanner used, design of experiment, etc. Therefore, the evaluation of preprocessing protocols with respect to their effectiveness or any errors they may have caused, will have to be done on a case by case basis. This is very much in line with the piecemeal approach of the error-statistical account as well as the essential notion that the severity of a test is always assessed postdata in terms of a specific hypothesis, a data set, and the experiment that generated the data set. When we place the severity function, SEV(T, $\mathbf{x}_0$, H) in the context of fMRI experiments, we can think of the preprocessing protocol as another aspect of T, that is, the experiment that generated the data. As Lazar (2008) and LaConte et al. (2003) seem to suggest, we can apply different preprocessing protocols to the same set of raw fMRI data and then do statistical analyses on the data sets that the different preprocessing protocols yield. In this way, we can assess the effects of these protocols on the same data set. The results of these analyses can be helpful in finding out how different preprocessing protocols affect the severity of the experiment as a whole as tests of hemodynamic real effect hypotheses. Informally speaking, in some experiments where the functional SNR is not very low extensive smoothing makes significant activations in some voxels highly probable even under the null hypothesis of no effect. Thus, extensive

smoothing may decrease the severity of an fMRI experiment in cases where the hypothesis of interest predicts high activation in a brain region. If the results of analyses of data show that there was high activation in that region, then the severity of this experiment would be lower than an experiment where the high activation is obtained from a data set with less smoothing. One crucial thing to note here is that as practitioners become more aware of the errors that preprocessing procedures may introduce, they start devising methods of identifying and controlling for the ways in which these errors arise in fMRI experiments. The NPAIRS framework is a nice example to this kind of work in the fMRI literature. The error-statistical notions of error probabilities and severe tests can aid this methodological trend by supplying additional criteria for the evaluation of inferences in fMRI studies.

4.3.2: Statistical Modeling of fMRI Data: As most other fields of scientific research, fMRI, too, makes use of statistical models and significance tests. As was identified in chapter three, statistical significance tests in fMRI studies test hypotheses about the relationship between a cognitive process and hemodynamic activity in the brain. In these tests, the null hypothesis states that there is no significant difference in the amount of hemodynamic activity, as measured by the fMRI scanner, between the experimental condition in which subjects perform a cognitive task and the control condition in which subjects do not perform any task or they perform a different task. The alternative hypothesis simply states that there is a significant difference in the amounts of hemodynamic activity between the experimental and the control conditions. Many fMRI experiments use significance tests, which take the voxel as their object of analysis.

Researchers employ correlation analyses, simple t-tests, or analysis of variance techniques on preprocessed fMRI data. For any of these tests to be applied without biases or errors, fMRI data sets have to be modeled carefully and adequately.

The general linear model (GLM) is commonly used in analyzing fMRI data (Huettel et al., 2004; Lazar, 2008). The factors in the GLM represent the hypothesized components of the data. Given the experimental data and model factors, one calculates the combination of factor weights that minimize the error term. If there is only one model factor, then the GLM is identical to a correlation analysis; if there is only one model factor with two levels, then the GLM is identical to a t-test. The form of the GLM can be expressed in the equation:

$$Y = X\beta + \varepsilon$$

where Y is the preprocessed fMRI data, which may be represented in a matrix of the time series data from all voxels, so it will have one column for each voxel and one row for each time point (Lazar, 2008; p.83). In the GLM equation, X represents the model factors, and it can be expressed in terms of a design matrix which represents what stimuli or tasks were presented to the subjects during the course of the experiment. For example, pictures that were shown to subjects and tasks they were asked to perform, and the time points at which these were presented would be included in the design matrix. β represents the unknown coefficients of the model factors and ε represents the error, which is assumed to be normally distributed with mean zero and variance σ^2 (ibid.). GLM in this form provides a basic example of how statistical tests are thought of in the fMRI literature. Statistical tests are conceived as tools to find out which experimental manipulations, i.e. factors, have the greatest effect on the preprocessed fMRI data. In

other words, statistical tests are designed to discover whether or not manipulations of cognitive tasks produce significant increases in hemodynamic activation in the brain as a whole, or certain regions of the brain.

As any other statistical model, the GLM comes with a set of probabilistic assumptions about the data generating mechanism. These are: 1) The data Y is normally distributed; 2) The process that generated the data Y is an independent process; 3) The expectation of data Y is linear in X; 4) The variance of data Y is homoscedastic, i.e. variance of Y is free of factors X (Spanos, 1998). Cognitive neuroscientists use the GLM to model fMRI data where they assume that: 1) Raw fMRI data can be modeled as the sum of separate factors and additive Gaussian noise, 2) Each factor may vary independently across voxels, and 3) Gaussian noise is independently and identically distributed (Huettel et al., 2004; p.342). Naturally, the verification of these assumptions is of crucial importance to establish the validity of statistical inferences. Petersson et al. reviewed several statistical methods and models used in functional neuroimaging research and they have concluded that "assessing model fit and verification of assumptions are challenging tasks and effective tools for assessing the goodness-of-fit of models and diagnostics for violations of assumptions are generally lacking in FNI [functional neuroimaging]" (1999; p. 1256). About a decade later, Lazar (2008) states that the assumptions of the GLM in the context of fMRI “are surely unrealistic and hence violated in practice…” (p.85). These two statements together are quite telling of how challenging are the problems of modeling fMRI data that practitioners and statisticians are still trying to find ways of addressing them.

Petersson et al. (1999) discuss the difficulty of modeling baseline fluctuations of the global activity of the brain. In an fMRI experiment, baseline fluctuations may be large enough to hide the task-related signal. For a chosen model to work, so that we can draw valid statistical inferences from modeled data, it needs to include a characterization of the global baseline activity of the brain and its fluctuations as accurately as possible, because regional activations are measured relative to this baseline activity. Many experiments take the fMRI data from control conditions, in which the subjects are at rest, as the baseline activity of the brain. The task-related signals are investigated relative to this baseline condition, that is, researchers look for voxels or regions in the brain that are significantly more active when subjects perform the cognitive task compared to baseline activity when the subjects are at rest. At first look, this seems to be a reasonable assumption about the brain, which is incorporated into the models of data. However, this simple assumption may introduce errors in the data modeling process. In a series of experiments, Stark and Squire (2001) used fMRI data from rest conditions as the baseline activity of the brain as a whole. They asked subjects to perform episodic memory encoding tasks in the experimental conditions as fMRI data were collected. The results showed that activity in the medial temporal lobe and some other regions were higher in the rest condition than in other conditions. This ended up reducing or totally masking task-related signals of interest. Thus, if researchers use fMRI data from the rest condition as the baseline activity level of the whole brain, this may cause some task-related signals to go undetected. One needs to check for this possibility; but doing this requires effective ways of combining substantive knowledge of normal activity patterns of the brain and techniques of modeling fMRI data. The question of modeling baseline activity of the

brain is an example among many problems that arise in model validation procedures. Unfortunately, as Lazar (2008) notes, although there is a sizable general statistical literature on model validation and model assessment, this literature has not been adapted and employed sufficiently in fMRI data analyses.

Lazar (2008) discusses two general questions that arise in model validation in fMRI. One is a question about which model, among many alternative models, should be chosen to fit to the brain as a whole. One problem here, as Lazar states, is that the notion of fit in this context does not have a precise definition. The other question is about whether or not the same model should be fit to every voxel in the brain. Given the variability of fMRI data across voxels, it appears that if the same model is fit to every voxel, some voxels will be underfit while others will be overfit. However, if different models are fit to different voxels, some necessary statistical procedures cannot be used. For example, detecting contiguous groups of active voxels is crucial for any fMRI experiment and one way in which this is accomplished is by applying random field thresholding. However, this thresholding technique cannot be used if different models are used for different voxels.

As serious as the above problems are, the fundamental problem in modeling fMRI data is the fact that assumptions of standard models such as GLM are violated in the practice of fMRI research. For example, the assumption of independence in the GLM is violated in fMRI studies. This is because regions of the brain are densely connected with each other and when one regions is activated this causes activations in nearby regions as well. Thus, the process that generates fMRI data is not always an independent process. This may threaten the validity and reliability of statistical inferences drawn from fMRI

data. This problem may be one of the factors responsible for the relatively high incidence of contradictory findings in the fMRI literature. Lazar (2008) calls attention to several drawbacks in fMRI analyses that are caused by problems of model validation. Some of these drawbacks are misspecification of models, choosing oversimplistic models due to a lack of criteria for systematic evaluation of models, and improper choice of models on the basis of number of active voxels, where a model is considered to be a better model if it detects more active voxels. This is not a very healthy method of model validation because, as researchers desire to be able to detect more voxels as active they may choose a model that is not statistically adequate which may introduce biases in the data analyses and lead to erroneous inferences.

In this environment where fundamental problems of data modeling threaten the validity and reliability of inferences, as recognized by fMRI researchers like Petersson and his colleagues and Lazar as well as others, the error-statistical approach to model validation proposed by Mayo and Spanos (2004; 2010; 2011) can be useful. One central aspect of this approach is misspecification (M-S) testing, which includes methods of testing the assumptions that models make about the data generating mechanism. Respecification is also an essential element of M-S testing; if assumptions of a model are violated, iterative procedures are applied to accommodate flawed assumptions in respecified models. In the end, a statistically adequate model of the data at hand is obtained, which can support reliable inferences about the hypotheses of interest. Another advantage of M-S testing is that it distinguishes between problems of model specification and problems of model selection where researchers select a model from an assumed family of models, which appears to be common practice in the fMRI literature. M-S

testing provides a method for developing statistically adequate models of given data sets. Once we have a data set, say preprocessed fMRI data, we can proceed by what Mayo and Spanos (2004) call the probabilistic reduction approach in which we think of the set of all possible statistical models of the mechanism that generated the data. Every statistical model is a set of probabilistic assumptions about the data generating mechanism. These assumptions can be grouped under three broad categories which are; distribution, dependence, and heterogeneity. Given a specific fMRI data set, we can start the specification process by asking questions about the data set, such as 'are the data from the same voxel independent over time?', 'are the data from different voxels independent?', 'what is the distribution of the data? e.g. normal or skewed?', 'are the data from different voxels, or different regions of interest, identically distributed?' The answers to these questions will eliminate certain possibilities for the model to be chosen. For example, as has been noted before, data from neighboring voxels are not independent. In fact, often there is spatial correlation between data from adjacent voxels as is to be expected given the highly connected anatomy and functioning of the brain. Thus, any model that cannot accommodate this dependence in the data would be eliminated as a potential model. Obviously, given the immense size and complexity of fMRI data sets, the application of M-S testing to fMRI data will be a serious undertaking. However, by proceeding within the probabilistic reduction approach, statistically adequate models of fMRI data can be developed, which would be a significant boost to improving statistical analyses of fMRI data.

4.3.3: Multiple Testing and Thresholding in fMRI: Statistical significance tests in fMRI test hypotheses about the relationship between performance of a cognitive task and hemodynamic activity in the brain. In these tests, the null hypothesis states that there is no significant difference between population means of hemodynamic activation between the conditions in which subjects perform a cognitive task and the control condition in which subjects are at rest or perform a task other than the task of interest; H_0: $\mu_1 - \mu_0 = 0$. The alternative hypothesis states that there is a significant difference between population means of hemodynamic activation across the experimental and the control conditions. In most cases, the alternative is a directional hypothesis that predicts significantly higher activation in certain brain regions in the experimental condition with high probability; H_1: $\mu_1 - \mu_0 > 0$. Most significance tests in fMRI studies take the voxel as the object of analysis. As Huettel et al. (2004; p.333) state; "The goal of most fMRI statistical tests, regardless of their complexity, is to evaluate the probability that each voxel is consistent with the null hypothesis." That is, the statistical test in fMRI is done to find out how probable certain outcomes are if the null hypothesis is true. If this probability is very low, that is, smaller than .025, then we can infer that the alternative hypothesis H_1: $\mu_1 - \mu_0 > 0$ is true.

Since fMRI experiments yield time series data from thousands of voxels, a typical fMRI experiment requires thousands of significance tests. The statistical parametric map (SPM) is one of the tools used by fMRI researchers, which color-codes all voxels within an fMRI image according to the result of a significance test. If the probability associated with the activity in a given voxel is below a threshold probability, or alpha value, then that voxel is labeled as active. For example, if the threshold alpha is set at .01 and a given

voxel is at .009, then it is displayed dark red; if the voxel is at .000001, it is displayed bright yellow (Huettel, 2004). The statistical parametric map, which is displayed on top of an anatomical image of the brain, reflects the outcomes of significance tests done for every voxel or group of voxels. Thus, the SPM represents the correspondence of the data to a hemodynamic hypothesis of interest. The SPM provides the final neuroimages that are presented as the fMRI findings to scientists and the public.

The fact that researchers have to do significance tests on thousands of voxels creates serious problems of multiple testing where many of these significance tests will yield significant results due to mere chance. That is, a lot of these significance tests would be committing the type I error and their results can be designated as false positives. The problem of multiple testing makes choosing an appropriate threshold for significance a critical step in making reliable inferences. This is a well-known problem arising in cases where a large number of significance tests have to be done. Citing the multiple testing problem, some philosophers of science have criticized psychologists and cognitive neuroscientists for using significance tests in such cases, e.g., Colin Klein has raised doubts about the reliability of fMRI findings because of this problem. According to Klein (2010b), choosing an overly liberal threshold for significance may lead to significant results in the absence of a real effect. That is, when we observe significant activation in a given group of voxels, this may be due to a liberal threshold for significance, which may pick up background noise as genuine task-related activation.

As multiple testing is a well-known problem, there are various ways of correcting for it in the statistical literature, some of which are adapted to and employed in fMRI studies. Lazar (2008) discusses several different methods of thresholding to correct for

multiple testing. In the early days of fMRI, a lot of psychologists were beginning to employ fMRI in their research. These psychologists started using the Bonferroni correction for familywise error rate on fMRI data, which is a common correction method for multiple testing in psychology. However, this proved to be an overly conservative method for fMRI data, because Bonferroni gave a corrected significance level which was defined as α/m, where α is the overall significance level, say .05, and m is the number of tests (ibid., p.188). Recall from above that in a typical fMRI study thousands of t-tests may have to be done. Thus, the Bonferroni corrected significance level, which was very low due to the great number of tests in fMRI, led to analyses that did not detect any active voxels. This was so even in cases where subjects were asked to do a task, which we know from neurobiological knowledge that leads to activation in certain regions of the brain, but analyses using Bonferroni corrected significance levels did not detect any active voxels in those regions of the brain (ibid.).

Because of the overly conservative nature of the Bonferroni method, new methods of correction for multiple testing had to be devised that could accommodate the large number of significance tests in fMRI. One of these methods is the cluster threshold method, which embodies the basic reasoning behind some other methods of thresholding. In these thresholding methods the major interest is in voxels that are declared as active voxels in significance tests due to mere chance, which may be designated as false positives. In other words, these are voxels for which the null hypothesis is rejected while it should not have been rejected. The aim of the cluster threshold method is minimizing the number of false positive voxels. The cluster threshold method, as well as thresholds obtained from random field theory or permutation, was motivated by a general

observation about brain anatomy and functioning (Lazar, 2008). Recall from above that data from neighboring voxels are correlated, that is, activation in a voxel influences nearby voxels to be more active, too. This is to be expected when we remember the highly connected anatomy of the brain and this is also related to how the independence assumption of the GLM is violated in fMRI studies. Let us also recall that the size and number of voxels are defined by the researcher, so voxels do not carry any physiological or anatomical significance, they are merely partitions of data that correspond to chunks of brain tissue. Thus, Forman et al. (1995; cited in Lazar, 2008) conjectured that task-related activations in the brain, that is, activations that are indeed caused by cognitive performance, are probably spread across several contiguous voxels. Also according to this conjecture, one isolated voxel detected as active are probably false discoveries, since the brain tissue corresponding to a single voxel probably does not give rise to a cognitive function by itself. By this reasoning, many researchers treat active single voxels as false discoveries. Lazar states that these conjectures are supported by simulation results (ibid., p.189).

In the cluster threshold method, researchers determine the probability of false positives by setting two criteria: 1) $C(\alpha)$: the significance level for rejecting the null hypothesis for a given voxel at α, and 2) S: the size of a contiguous cluster of voxels (Lazar, 2008; p.189). Thus, there are two conditions for a voxel to be declared active: first, the voxel must cross $C(\alpha)$ and second, a sufficient number of its neighboring voxels must also cross $C(\alpha)$. The number of neighboring voxels that must also cross $C(\alpha)$ is determined by S, the size of the cluster. So, the size of the cluster determines the smallest cluster that can be detected as active. That is, if a group of voxels are found to be active

but the size of the cluster they make up is smaller than the size set by *S*, then that group is not considered as an active group of voxels. So defined, researchers can try different combined thresholds by varying the values of *C(α)* and *S*. The advantage of this method is that researchers can calculate specific error probabilities, that is, probabilities of false positives associated with applying different thresholds, which come from various combinations of values of *C(α)* and *S*. For example, if *C(α)* is fixed, then increasing cluster size *S* has the positive effect of decreasing the probability of false positives. Therefore, error probabilities associated with different cluster thresholds would be part of severity assessments of fMRI experiments as tests of hemodynamic hypotheses. For example, if, in an fMRI experiment, a cluster is set to be too small with a fairly liberal significance level, then this may increase the probability of false positives. The results that agree with the hemodynamic hypothesis of interest that come from this experiment may have been obtained because of the size of the cluster and not because the hypothesis is true. Thus, this experiment would not constitute a severe test of the hemodynamic real effect hypothesis framed as H_1: $\mu_1 - \mu_0 > 0$. So, these results would not provide evidence for this hypothesis. This illustrates how the error probabilities associated with cluster thresholds can be incorporated in severity assessments of fMRI experiments. The results of these assessments can tell us whether or not we can reliably draw inferences to the truth of hemodynamic hypotheses of interest. This is also another example of the general notion of how we can learn what we can learn from fMRI experiments.

In this section, I discussed three different aspects of statistical analyses of fMRI data, namely the preprocessing procedure of spatial filtering, statistical modeling of data, and

thresholding to correct for the multiple testing problem. I focused on the error characteristics associated with these aspects, and where possible, I emphasized how actual error probabilities can be calculated as in the case of cluster thresholds. In terms of the hierarchical framework of models of inquiry, I identified three aspects of fMRI research, located them in the data models, and analyzed how they may influence experimental outcomes and lead to errors, and how these errors can be controlled (see figure 4.2).

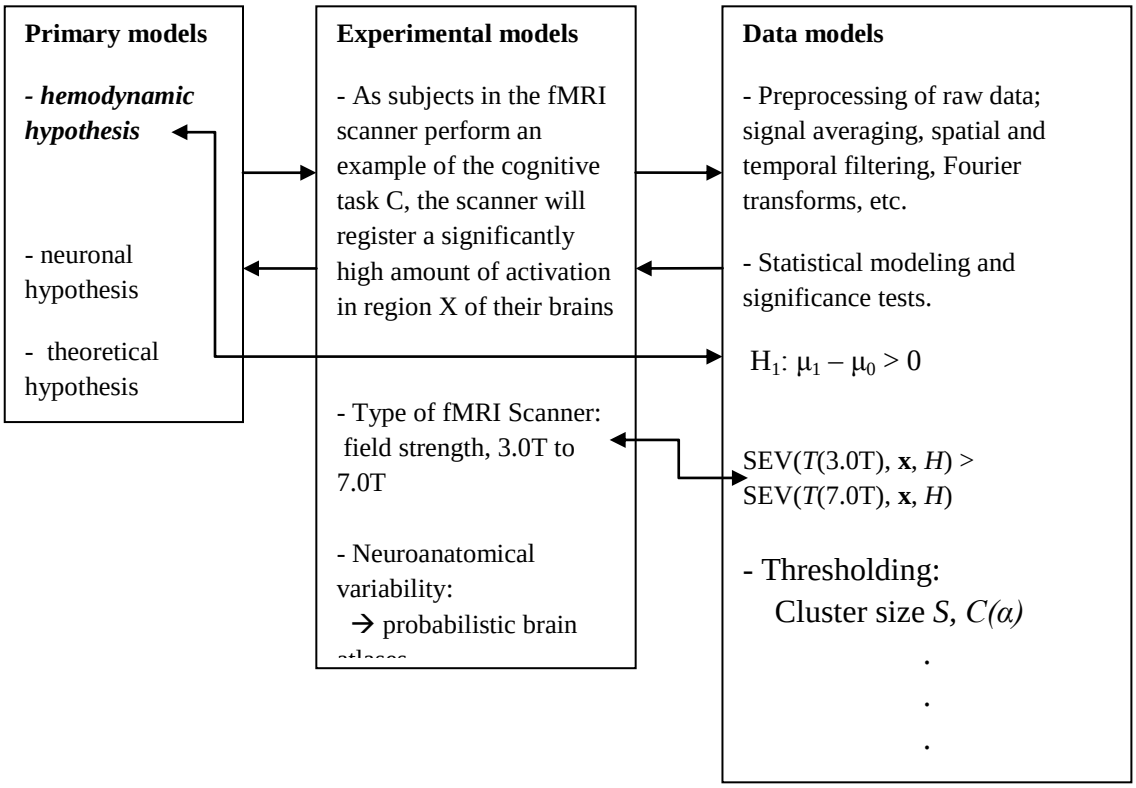

Figure 4.2: Data Models with Spatial Filtering, Modeling, and Thresholding As Sources of Error

In all three examples, the main theme has been the consideration of error characteristics or error probabilities associated with these aspects of fMRI. Again we have seen that the study of effects of these aspects on experimental outcomes is a focus of much

contemporary work in the fMRI literature. The NPAIRS framework, in which preprocessing protocols have been scrutinized with respect to their effectiveness, is an example of this kind of work. In this section, I also discussed how these aspects of fMRI can be thought of as parts of the data generating mechanism, or the test *T* in the severity function SEV(*T*, $\mathbf{x}_0$, *H*). As such, they can be thought of as factors that determine the severity of experiments as tests of hemodynamic real effect hypotheses framed in terms of statistical parameters, H_1: $\mu_1 - \mu_0 > 0$. Therefore, as experimental knowledge accumulates about the data models of fMRI, we will have a better and more complete understanding of statistical analyses and inferences in fMRI. This understanding can then be utilized in elucidating how we can learn from fMRI studies and the nature of experimental knowledge in cognitive neuroscience.

4.4: Voodoo Correlations Revisited

In chapter one, I discussed in detail the voodoo correlations debate in social neuroscience. This debate came out of the work of cognitive neuroscientists Vul and his colleagues who looked at some correlation coefficients, which seemed to be too high to be true, reported in peer-reviewed scientific journals by fMRI researchers. Vul and his colleagues (2009) estimated the reliability of tests used in these studies as well as the reliability of fMRI as a research tool. On the bases of these reliability estimates and the assumption that the fMRI studies were done with no measurement errors, Vul and colleagues estimated an upper limit for possible meaningful correlations between scores on psychological tests and fMRI data. This upper limit was calculated to be .74. Let us note again that this is the value we would get in the ideal case, where the reliabilities of

psychological tests and the fMRI procedure are at the highest possible values and the underlying correlation between the measured trait and the pattern of brain activity associated with the trait is perfect. In actual studies, this state of affairs is not probable. However, the puzzling thing about these reported correlations in social neuroscience journals was that a lot of these correlations exceeded the upper limit of .74. If the upper limit for any meaningful correlation is .74, then what do these correlation coefficients higher than .74 mean? Vul and his colleagues have focused on how these overly high correlations were calculated. They sent surveys to the authors of these articles regarding the methods used. In these surveys, the critical question was how the voxels were selected the fMRI data from which were correlated with scores on psychological tests (ibid.). Were they selected on the basis of anatomical or functional criteria, or both? If functional data were used to select the voxels, were they the same functional data as were used to define the reported correlation?

The results of the survey showed that 53 percent of the respondents said that "regression across subjects" was used as a functional constraint to select voxels "indicating that voxels were selected because they correlated highly with the behavioral measure of interest" (Vul et al., 2009; p. 278). Furthermore, it was found in the survey that all these studies used "the same data to compute the correlation as were initially used to select the subset of voxels" (ibid.). The procedure was like this: researchers give each subject a psychological test and they obtain a behavioral measure from each subject. They also collect fMRI data as subjects do the test. Then, the fMRI activity observed in each voxel is correlated with the behavioral measures, which yields thousands of correlation coefficients. After this, the voxels for which the correlation passes a statistical

threshold are selected and the fMRI signal is aggregated across the selected voxels and a final correlation coefficient is calculated between the fMRI signal and subjects' scores on the behavioral measure. Vul and colleagues state: "Such an analysis will inflate observed across-subject correlations, and can even produce significant measures out of pure noise. … With enough voxels, such a biased analysis is guaranteed to produce high correlations even if none are truly present" (ibid., p. 279). They identify the methodological flaw in this procedure as the "non-independence error," which is selecting one or more voxels based on a functional analysis, and then reporting the results of the same analysis and functional data from only the selected voxels. They write: "This kind of analysis distorts the results by selecting noise exhibiting the effect being searched for, and any measures obtained from such a nonindependent analysis are biased and untrustworthy" (ibid.).

What Vul and colleagues call the non-independence error has also been called "double-dipping" by other authors. This name reminds one of what Deborah Mayo calls "double-counting of data" (Mayo, 1996; Mayo, 2008). This concerns: if data **x** are used to construct a hypothesis $H(x)$, then should the same data be used again as evidence supporting $H(x)$? Those who say 'no' generally do so because they want to prevent data mining or hunting for significance which could lead to unreliable inferences.

At first look, such concerns may appear to be relevant to the issue Vul et al. raise. However, in the error-statistical account, one does not make a blanket statement about double-counting and prohibit all instances of it. Rather "it is the severity, stringency, or probativeness of the test—or lack of it—… that should determine if a double use of data is admissible…" (Mayo, 2008; p.858). Thus, in a given experimental context one asks the question: does this type of double-counting influence the error probabilities of this

experiment in testing the given hypothesis? So, it is precisely this question that we must ask about what Vul et al. call the nonindependence error, that is, we need to know whether it is double-counting or some other aspect of the experiment that yields the voodoo correlations. But before asking this question we should ask, what kind(s) of double-counting occur in the experiments Vul et al. discuss? The researchers in those experiments collect fMRI data from thousands of voxels and correlate the fMRI data with scores on psychological tests. It looks like double-counting occurs when the researchers select those voxels for which the correlations pass a statistical threshold. Now, if they stopped here and just reported the correlations for each of the selected voxels separately and state that these correlations constitute evidence in support of the hypothesis that these voxels are related to the behavioral measure, this would be an instance of double-counting. Yet, if every procedure of the fMRI experiment were run without serious flaws and the thresholds were chosen in ways that minimize false positives, then this instance of double-counting of data would not have greatly influenced the error probabilities of the experiment. Thus, not all instances of double-counting are bad. However, the problem, as Vul et al. say, arises when the mean of the fMRI data from the selected voxels is used in obtaining the final correlation value between fMRI data and the behavioral measure. Regardless of whether or not this is a case of double-counting of data, it appears that this procedure greatly increases the error probabilities of the fMRI experiment. The central question that the error-statistician would ask is 'what is the probability that this procedure would yield a high correlation between fMRI data and the given behavioral measure, if in fact this correlation is very low or zero?' The answer to this question is "very high!" Thus, there indeed is a problem but it is not double-counting

of data per se rather it is the increased error probability associated with the inference that there is genuine correlation.

The central issue in the experiments Vul and his colleagues discuss is that the error probabilities associated with the experiments that commit the 'nonindependence error' are too high. In other words, this kind of experiment would yield significant results with high probability, that is, it would yield significant correlations even if there is no real relation between fMRI data and scores on the given behavioral measure. Thus, Vul and his colleagues are right in saying that the correlations coming from fMRI experiments that used this procedure do not constitute good evidence for the hypotheses that they were designed to test. The real problem here is not double-counting but it is the fact that experiments that use the procedure Vul and his colleagues describe are low severity tests of the inference that there is a genuine correlation between behavioral measures and fMRI data.

Recall the dead salmon demonstration discussed in chapter one where a dead salmon's brain was found to be active when it was shown emotional faces and asked to identify the emotion on the faces. If an fMRI experiment does no better than the salmon experiment in identifying and controlling for errors, then the data it yields are certainly no evidence for any hypothesis. The error-statistical analyses applied in this chapter enable us to find out whether an experiment does a good job in controlling for errors. These error-statistical analyses into error probabilities of component parts of fMRI studies can help us identify the conditions under which we can draw reliable inferences from fMRI data to hemodynamic hypotheses about relationships between cognition and brain activity. This is how we can learn from fMRI.

Chapter Five:
Experimental Knowledge and Progress in Cognitive Neuroscience

This dissertation is a work in the epistemology of functional neuroimaging (fNI), the major research paradigm of cognitive neuroscience. The field of philosophical and methodological study of fNI can be thought of as a subfield within philosophy of psychology, a subfield that has not been sufficiently developed. There are two reasons for this; one reason, as Machery (forthcoming) notes, is that philosophers of psychology have rarely discussed issues of evidence and inference in psychological science. Another and more obvious reason is that fNI is still a new and developing paradigm, so it is natural that only a few philosophers of science and a few practitioners have turned their epistemological gaze toward fNI research. However, these few authors did provide a literature on the epistemology of fNI. One of the objectives of my work here was to contribute to the epistemology of fNI by applying the error-statistical (ES) philosophy to issues and problems of fNI and find out how the ES account can help us formulate and address these problems. I aimed to help the further development of the epistemology of fNI to give us a clear, accurate, and more complete understanding of what we can learn from fNI and how we can learn it. Now, I will discuss the implications of the results of my work for fNI as a developing research paradigm, and especially for the growth of experimental knowledge in cognitive neuroscience. But first, let me briefly summarize the contents of chapters one through four.

5.1: A Brief Recap

In chapter one, I grouped the works in the epistemology of fNI into two categories; the first category consisted of discussions of the theoretical significance of fNI findings and the second category consisted of discussions of the methodological difficulties of fNI research. The major focus of the discussions in the first category was the question of what we can learn from fNI results and how, if at all, these results can be used in evaluating theories of human cognition. If fNI results can be reliably established, then they can be used in adjudicating between rival theories of cognition, where if a certain theory conflicts with fNI findings then it is revised or rejected. This kind of work mostly ignored essential methodological aspects of fNI research, even though the discussion of these aspects is necessary for any assessment of the theoretical significance of fNI findings. The work in the second category emphasized methodological issues which mostly had to do with the high degree of complexity of the workings of fNI tools, the immensely large data sets that fNI yields, and the difficulty of statistically modeling these large data sets and obtaining reliable inferences.

I identified some features of the literature, which I saw as shortcomings. One was that the work in the first category has been too theory-centered in its approach. The major concern in these works was whether or not fNI findings can be used to evaluate large-scale theories of human cognition. This theory-centered approach is not very useful for several reasons. Constraining the epistemology of fNI to a theory-centered approach precludes the possibility of gaining important insights into general questions regarding the kind of knowledge we *can* and *do* gain from fNI studies. For example, the authors who are skeptical about the theoretical significance of fNI findings have raised the

problem of theory-ladenness as another reason for their skepticism. Both Uttal (2001) and Hardcastle and Stewart (2002) have argued that prior to experiments fNI researchers assume the existence of specific cognitive processes, which are localized in different parts of the brain, and this creates a serious problem of circularity in fNI research. So, these authors concluded that fNI findings are hopelessly theory-laden in the terms of modularist theories of cognition. In chapter three, I discussed how we can avoid problems of this kind of theory-ladenness. I showed how fNI results can provide support for non-modularist hypotheses about cognition, namely hemodynamic hypotheses about relationships between cerebral blood flow and performance of cognitive tasks, and on the basis of this result I discussed how we can talk fruitfully about fNI findings independently of modularist or non-modularist theories of human cognition.

Another shortcoming of the theory-centered approach is the lack of emphasis on methodological aspects of fNI. Yet, problems about the theoretical significance of fNI findings cannot be resolved without adequate scrutiny of the methodological characteristics of fNI tools. Some authors put forth skeptical arguments the conclusion of which was that fNI findings were theoretically irrelevant or useless. One premise for these arguments was that inferences drawn from fNI data depended on unreliable procedures. But, of course, this is a methodological problem and any account that addresses the epistemic value of fNI findings is required to address such methodological issues. One way to address skepticism about the epistemic value of fNI data is by going into the characteristics of the methodology of fNI. Analyses of the methodologies of fNI can not only give us ways to address skepticism but also new insights into the nature of experimental knowledge in cognitive neuroscience as well as into general philosophical

problems such as theory-ladenness. I demonstrated this in chapter two, where I discussed the history of the development of fMRI. I showed how the workings of fMRI give rise to a kind of theory-ladenness that is useful for cognitive neuroscience and different from the kind discussed by the skeptics. Since the fMRI scanner works on the bases of physical and physiological theories, findings of fMRI studies are laden in the theories of physics and physiology. Because these theoretical bases have been well-established by experimental knowledge that physics and physiology have produced, we know what kinds of errors they may introduce errors and we can control for these errors. Thus, this kind of useful theory-ladenness allows for the representation of and intervention in constructs of cognitive neuroscience using a well-tested, well-established tool of research namely the fMRI scanner. This provides additional support for the reality and manipulability of these constructs independently of modularist theories of cognition. It also enables us to check whether or not fMRI experiments constitute severe tests of hemodynamic real effect hypotheses. In a sense, this kind of useful theory-ladenness makes it possible to do psychology from the outside.

In chapter one, I also discussed the part of the philosophical literature on fNI that discusses specific methodological problems that arise in fNI research. Although the emphasis on methodological details is definitely correct, there is a general assumption in these works that the methodological difficulties of fNI cannot be satisfactorily addressed. This assumption was explicitly or implicitly expressed in the works of Klein (2010b) and Roskies (2008; 2010). For example, Klein has argued that problems in the use of statistical significance tests in fNI cannot be satisfactorily resolved and Roskies argued that the great number of necessary preprocessing procedures on fNI data lower the

reliability of inferences in an intractable way. On the basis of these arguments, both Klein and Roskies have raised what can be thought of as Duhemian challenges and have concluded that hypotheses and theories about human cognition are generally underdetermined by fNI data. I showed in chapters two, three, and four how problems of underdetermination can be addressed as they arise in fMRI.

I addressed skeptical arguments against fMRI in chapters three and four, where I applied the error-statistical notions of error probabilities and severity of tests together with the hierarchical framework of models of experimental inquiry. In chapter three, I discussed the problem of underdetermination as it arises in fMRI. I asked the question 'what kinds of hypotheses can be put to severe tests in fMRI experiments?' In terms of the hierarchical framework of models of inquiry, this question can be thought of as asking what can legitimately be placed in the primary models of an fMRI study. By looking at how fMRI works and the kind of measurements it gives us, I showed that one kind of hypothesis, namely hemodynamic hypotheses, need not be underdetermined by fMRI data. This is because fMRI provides data on hemodynamic processes in the brain as subjects perform cognitive tasks and hemodynamic hypotheses can be put to severe tests in fMRI experiments. I also discussed in chapter three how hemodynamic hypotheses can be theoretically significant and how they can fuel the growth of experimental knowledge in cognitive neuroscience. Hemodynamic findings from fMRI experiments can be used in constructing new theories as well as evaluating existing theories of cognition where any current theory that conflicts with a well-established hemodynamic finding would have to be revised. Thus, I showed in chapter three what it is that we can learn from fMRI.

In chapter four, I addressed the Duhemian challenges raised by the skeptics. Continuing to use the hierarchical framework of models of inquiry, armed with the ES notions of error probabilities, or error characteristics, I addressed the question 'what can be placed in the experimental and data models of an fMRI experiment?' Any fMRI experiment is a highly complex undertaking that involves several different procedures from data collection to preprocessing and modeling and analyses of data. If we know how these aspects and procedures influence the results of fMRI experiments independently of the truth or falsity of hypotheses, then we can find out on a case by case basis whether or not given experiments constitute severe tests of the hypotheses they were meant to test. I showed in chapter four how different aspects and procedures of an fMRI experiment can be placed in the experimental or data models. Once properly located, I discussed how each aspect or procedure may influence the results of the experiment and lead to errors. For example, I argued how fMRI scanners of high magnetic field strength may lead to false discoveries by detecting noise as real effects because of their high sensitivity. I also showed how different procedures of statistical thresholding, placed in data models, can lead to type I errors and how error probabilities associated with different set thresholds can be calculated. These discussions gave us the error probabilities or error characteristics associated with these procedures, which can then be incorporated into severity assessments of fMRI experiments.

Throughout this dissertation, I adopted an approach that puts the emphasis on the experimental knowledge we obtain from fNI studies. I took fMRI as my case as I discussed the problems of evidence and inference that have been raised about fNI research. I employed the error-statistical philosophy (Mayo, 1996; 2005a; Mayo &

Spanos, 2010; 2011) to formulate and address these problems. This work as a whole offers an error-statistical epistemology of fNI, which hopefully will be a fruitful contribution to the philosophy of cognitive neuroscience. Now, I wish to discuss the implications of my work for cognitive neuroscience, especially about the nature of experimental knowledge in this developing field.

5.2: Experimental Knowledge and Progress in Cognitive Neuroscience

It is natural to expect of any epistemological study of a field to give us clear statements on the kind of knowledge we can obtain from that field. I attempted to do this for cognitive neuroscience by focusing on the kind of experimental knowledge we can reliably obtain from fNI studies. The work in chapters three and four show that fNI studies, aided by sophisticated computational and statistical techniques, can provide cognitive neuroscience with two kinds of experimental knowledge: one has to do with knowledge of how instruments and procedures of fNI work, discussed in chapter four, and the other has to do with knowledge of relationships between hemodynamic processes in the brain and cognitive processes, discussed in chapter three. Both kinds of knowledge fuel the progress of cognitive neuroscience.

5.2.1: Experimental Knowledge of Instruments and Procedures: The short history of fNI has had a fast and constant stream of development and improvement of its tools and techniques. This can be thought of as the history of growth of experimental knowledge about the tools and techniques of cognitive neuroscience. For example, early cognitive neuroscience got its start with the research technique of electroencephalography

(EEG), which had good temporal resolution but poor spatial resolution, and then with positron emission tomography (PET), which had better spatial resolution and unsatisfactory temporal resolution. However, both techniques have been fruitfully used with EEG being still in use though in combination with other techniques. Then, as we have seen in chapter two, fMRI arrived in the 1990's, which has good spatial resolution and satisfactory temporal resolution. So, it is no surprise that beginning in the mid 1990's it has dominated cognitive neuroscience and led to an explosion of fNI research. With fMRI at hand as a major technique, practitioners started studying and improving on certain aspects of fMRI as well as issues related to the use of fMRI. For example, the development of probabilistic brain atlases, discussed in chapter four, section 4.2.2, is an ongoing research stream where practitioners devise ways of addressing experimental flaws and errors that may be caused by neuroanatomical variability across subjects. The progress from static Talairach and Tournoux atlas to probabilistic brain atlases illustrates how practitioners identify certain aspects of a research paradigm and the errors those aspects may cause. To control for these errors they think of new techniques, they devise procedures to carry out these new techniques, and they test whether or not their procedures work properly to control for the identified errors. This is an example of the growth of experimental knowledge of instruments and procedures in cognitive neuroscience.

Another example can be found in chapter four, section 4.3.1 where I discussed preprocessing procedures in fMRI and the NPAIRS framework. This framework is a set of statistical analyses devised by Strother and colleagues to test for the effectiveness of different preprocessing protocols which include different preprocessing procedures of

varying degrees. For example, one protocol includes temporal filtering of raw data while others do not. Or, different protocols have varying degrees of spatial filtering of raw data. The NPAIRS framework allows researchers to find out which preprocessing protocols are more effective in improving signal-to-noise ratios (SNR) and controlling for certain errors. This, too, is a new trend in fMRI research, which arose out of the problem that preprocessing protocols may introduce false discoveries. Also, in some cases these protocols fail to improve SNR and yield experimental tests that lack sufficient power to detect any effects of interest. The NPAIRS framework grew out of the real need to address these methodological problems. Thus, as researchers engage in fMRI research they discover how things may go wrong and come up with methods to deal with them in order to do better experiments.

There are several other examples to this kind of experimental knowledge of instruments and procedures, such as devising new ways to design fMRI experiments, new ways of analyzing fMRI data. Some examples point to future directions in the development of methodological knowledge one of which I discussed in chapter four in section 4.2.1 where I suggested that the use of fMRI scanners of different magnetic field strengths should be scrutinized with respect to their error characteristics. Researchers may use scanners of field strength from 1.5 to 7.0 Tesla and the higher the field strength the more sensitive the scanner. The use of more sensitive scanners increases SNR considerably, which provides better power of detection. But at the same time more noise is detected by highly sensitive scanners, which may increase the chances of detecting noise as real effects. So, we need to find out how often scanners of certain strengths detect noise as real effects, or in other words how often they yield false positives. This is

nothing but the study of error characteristics of scanners and with sufficient study we can eventually have frequentist error probabilities associated with false positive rates for each type of scanner. To the best of my knowledge, this project has not been sufficiently pursued, yet it is certainly necessary. In sum, all these examples point to the importance of research and development of methodological techniques, which provide experiments with better designs, better measurement instruments, and better procedures of analyses. This is the growth of experimental knowledge of instruments and procedures and it has to be at the center stage to ensure the progress of cognitive neuroscience as a field.

5.2.2: Experimental Knowledge of Hemodynamic Substrates of Cognition: The other kind of experimental knowledge that fNI can provide is knowledge of the relationships between hemodynamic processes in the human brain and performance of cognitive tasks. After all, our thirst for knowledge of neural substrates of cognition is what fueled the birth and development of cognitive neuroscience in the first place. In chapter three, I showed in detail how fMRI experiments can put hemodynamic hypotheses to severe tests and how knowledge of the truth or falsity of hemodynamic hypotheses can be theoretically significant. Here, I wish to illustrate this with a well-known set of hemodynamic hypotheses that were proposed on the basis of data from the early days of cognitive neuroscience.

This set of hemodynamic hypotheses is known as the hemispheric encoding/retrieval asymmetry (HERA) model, which was proposed by Endel Tulving and his colleagues in 1994 (Tulving, et al, 1994). When it was first introduced, the HERA model was a straightforward description of empirical regularities found in PET studies of

memory (in 1994 fMRI was new and PET was the most common technique in cognitive neuroscience). The empirical regularities had to do with differential activation patterns in left and right prefrontal cortical regions when subjects engaged in encoding and retrieval tasks of episodic memories. In the commonly accepted view of human memory systems, episodic memory is one of two divisions of declarative memory. Episodic memory as a whole includes events, people, places, etc. that are personally experienced by an individual. The term autobiographical memory is sometimes used to describe this kind of memory; everything that is experienced by a person which he or she can remember would be stored in that person's episodic memory. For example, my first day in school is in my episodic memory; when I retrieve it, I remember that it was a sunny September day in 1983 and I was wearing a sweater with green and yellow stripes. The other division of declarative memory is semantic memory, which includes general knowledge of things that are not related to any specific personal experiences. For example, if you ask me the question 'who coined the term "pragmaticism"?', I can retrieve from my semantic memory that it was Charles Sanders Peirce who introduced this word to describe his philosophy and distance himself from other pragmatist philosophers. Yet, I have no recollection of the personal experience of or the physical context of obtaining this knowledge. Any contemporary textbook on cognitive psychology or cognitive neuroscience includes discussions on the external validity of these memory systems. These discussions are not relevant to my purposes, because here I am interested in observed relationships between hemodynamic processes and performance of cognitive tasks, which are operationally well-defined examples that can be classified as tasks of semantic or episodic memory (you can see section 3.4 of chapter three for a discussion of

operationally well-defined cognitive tasks as real processes). In the context of HERA, the terms ‘semantic memory’ and ‘episodic memory’ can be used simply to refer to these different cognitive tasks without having to worry too much about the theoretical baggage they may carry. In the experiments done or reviewed by Tulving and colleagues, the researchers asked subjects to perform tasks of semantic memory retrieval, episodic memory encoding, and episodic memory retrieval. The term ‘encoding’ refers to the processes of making a new memory, which may then be stored in the mind/brain. The term ‘retrieval’ refers to the process of recalling information from stored memories. These are terms that denote daily phenomena of memory that everyone knows and understands and they need not mean anything more theoretically complex than that in the context of the HERA model.

In 1994, when the HERA model was proposed, researchers did not know much about the neural substrates of cognitive processes in healthy humans. Cognitive neuroscience was just becoming a discipline of its own and new empirical regularities were starting to be observed in fNI experiments. In this environment of early development, the HERA model, which consists of a set of hemodynamic hypotheses about encoding and retrieval in episodic and semantic memory, was proposed in a data-driven manner on the basis of PET findings (although it works differently from fMRI, PET also provides measurements of cerebral blood flow, so PET data can be taken as evidence for hemodynamic hypotheses). Tulving and colleagues (1994) did a series of PET experiments and reviewed PET studies from different labs. In these experiments, researchers asked subjects to perform certain cognitive tasks; a common task was the verb generation task where subjects see or hear a noun, and are asked to produce a verb

that is related to that noun. For example, they see or hear the noun ‘ladder’ and they say ‘climb’. This is a task of semantic memory retrieval, because the related verb has to be retrieved from semantic memory of the subjects’ knowledge of the English language. But, as Tulving et al. (1994) state, this is also a task of episodic memory encoding, because the subjects encode the experience of saying the related verb into their episodic memory. Tulving and his colleagues also reviewed PET experiments in which subjects were asked to do explicitly defined episodic memory encoding tasks. For example, in one of these experiments subjects were asked to learn category-instance pairs, such as “poet-Browning”, and they were asked to recall these pairs later. This is an episodic memory encoding task since the subjects were told that they were going to be asked to recall the learned pairs and so they would encode the learning of these pairs into their episodic memory. In the experiments reviewed, subjects were also asked to do episodic memory retrieval tasks. For example, in one experiment, subjects were first shown drawings of objects or pictures of faces in the study phase. Later in the test phase they were again shown novel as well as previously studied drawings or faces and were asked if they recognized any of the drawings or faces from the study phase. In all these experiments, PET data were collected as subjects performed these cognitive tasks.

Thus, in these PET experiments subjects performed three types of cognitive tasks; tasks of semantic memory retrieval, tasks of episodic memory encoding, and tasks of episodic memory retrieval. When Tulving and his colleagues looked at the findings of these experiments, they saw some regularities in the observed patterns of brain activation as measured by PET. So, this was an example of exploratory research typical especially in the earlier days of cognitive neuroscience. They summarized these regularities in

statements, which are nothing but hemodynamic statements about the relationships between patterns of cerebral blood flow and cognitive processes of memory encoding and retrieval. These statements are: 1) Left prefrontal cortical regions are involved in semantic memory retrieval to a greater extent than right prefrontal cortical regions; 2) Left prefrontal cortical regions are involved in episodic memory encoding to a greater extent than right prefrontal cortical regions; and, 3) Right prefrontal cortical regions are involved in episodic memory retrieval to a greater extent than left prefrontal cortical regions. Essentially, these three hemodynamic statements constitute the HERA model.

The majority of the PET studies reviewed by Tulving and colleagues (1994) exhibited the regularities stated above with only a few exceptions, in which, for example high degrees of left prefrontal activation were observed in an episodic memory retrieval task. In 1996, Nyberg, Cabeza, and Tulving published another review article reporting results from both PET and fMRI experiments. Again, the great majority of these experiments exhibited the same findings predicted by the HERA model with only a few exceptions. Tulving and colleagues offered certain tentative explanations for the exceptions, i.e., experiments that did not obtain the findings predicted by HERA, by referring to flaws or differences in experimental design and tasks chosen for control and experimental groups. They argued that the chosen tasks were not sufficiently different to trigger high differential demand for retrieval or encoding activity to yield the pattern of activation found in the majority of experiments. In other words, these exceptions were not experiments that could have detected any differential patterns of activation as predicted by HERA. This review article also showed that the HERA model held not only for tasks with verbal stimuli but also for tasks that used nonverbal stimuli such as

drawings or photographs. However, Gabrieli and his colleagues (1998) used similar encoding tasks in which subjects were shown pictures and they found a significantly high degree of activation in the right inferior frontal cortical regions as subjects performed encoding tasks. In another study, Kelley and colleagues (1998) used an encoding task in which subjects were asked to study drawings of nameable objects and they found, like Gabrieli et al., a significantly high degree of activation in the right prefrontal cortical regions. Both sets of findings from Gabrieli et al. and Kelley et al. conflict with the HERA model, which predicts higher degree of activation in the left prefrontal cortex in encoding tasks. Some researchers concluded on the basis of these findings that the asymmetry of prefrontal cortex activations is due to the type of stimuli used, that is, verbal versus non-verbal, rather than being caused by the cognitive process involved, that is, encoding versus retrieval.

Unsatisfied with how their tentative explanations of findings, that do not agree with the HERA model, may be taken as ad hoc saves of HERA, Habib, Nyberg, and Tulving (2003) published an article in which they reformulated HERA to be stricter and more precise in its assertions. At the same time they insisted that HERA was a set of statistical hypotheses that compared degrees of activation in the prefrontal cortical regions between encoding and retrieval tasks rather than a set of absolute hypotheses which predict no activation in the left prefrontal cortex during retrieval and no activation in the right prefrontal cortex during encoding tasks. Tulving and colleagues offered a reformulation that includes two specific hemodynamic hypotheses which were expressed using abbreviations: ‘Enc’ meant encoding, ‘Ret’ meant retrieval, ‘L’ stood for a given left prefrontal cortical region and ‘R’ stood for the corresponding region in the right

prefrontal cortex. Combinations of task (Enc or Ret) and regions (L or R) stood for the observed activation during a given task in a given region, for example 'Enc L' stood for the activation observed in a specific region in the left prefrontal cortex during an encoding task. Thus, the two hemodynamic hypotheses that constitute the HERA model were stated: 1. (Enc L–Ret L) > (Enc R–Ret R); and 2. (Ret R–Enc R) > (Ret L–Enc L) (Habib, Nyberg, & Tulving, 2003; p. 241). These two formulations retained the comparative statistical character of the HERA model and added to it an important condition. This condition is that if one wants to test HERA by calculating the difference in activation between left and right cortical regions, one has to compare activations from specifically retrieval and encoding tasks. That is, as a control condition for a retrieval task one has to look at an encoding task and vice versa. The reason for this is that other control conditions such as being at rest or fixating at a cross on the screen cannot distinguish adequately between encoding and retrieval processes. On the basis of the new formulation of HERA, Tulving and colleagues argue that the findings from the Gabrieli et al. and Kelley et al. studies described above do not count as evidence against HERA, because in these studies encoding and retrieval tasks were not systematically manipulated and so brain activations were compared across encoding and retrieval tasks. The HERA model was reformulated in 2003. Since then several fMRI experiments, as well as experiments that used newer fNI techniques, yielded results that support the HERA model (for example, see: Babiloni et al., 2006; Cole, 2006; Thimm et al., 2010; and Okamoto et al., 2011).

There are several points to note about the development of the HERA model. The first thing to note is that it was proposed as a description of a set of hemodynamic

findings showing an asymmetry between encoding and retrieval tasks in episodic memory. Interestingly, some strict interpretations of modularist theories imply that there would not be a HERA as a single model. For some of these theories, fNI findings make sense only if they talk about cognitive modules executing highly specific computations defined by the theory and located in a neuroanatomically distinct part of the brain. This is why some authors think that before we can use fNI techniques to study human cognition we first have to have an accurate and complete psychological theory of human cognitive systems down to the tiniest separate cognitive function. Yet, all the experiments done or cited by Tulving and colleagues reported activations from different parts of the prefrontal cortex. These parts were all in the prefrontal cortex but were adjacent to each other or a few centimeters apart. This may be due to individual differences, that is, my memory of a certain face is located probably in a different place in my prefrontal cortex than where your memory of the same face is in your prefrontal cortex. For some modularist theories different parts of the prefrontal cortex may not be neuroanatomically distinct enough, or at least it is not clear whether or not they would be distinct. Thus, the HERA model will not make sense for anyone who adheres to such modularist theories. For some, it may constitute a refutation of these strictly modularist theories of cognition. Nonetheless, the HERA model still stands as a set of hemodynamic findings, because the prefrontal cortex is different enough from other parts of the brain and it certainly makes sense to talk about an established empirical regularity in the observed brain activations in the prefrontal cortex as subjects perform encoding and retrieval tasks. In addition, HERA came out of a data-driven approach and was described mostly in terms of hemodynamic hypotheses, which did not assume too much about cognitive psychological theories of human memory

systems. For example, take away terms like semantic memory or episodic memory with all their meanings in the context of large-scale theories of memory; we can still talk about the HERA model in terms of specific, well-defined memory tasks. The HERA findings would still stand even if theories of cognitive psychology change and we have a radically different way of dividing and categorizing human memory. The HERA model and the fNI findings out of which it came would endure as theories of cognitive psychology change.

After the HERA model was proposed, several groups of researchers have done various experiments testing it and a great majority of these experiments yielded results that support HERA. Also, however, it was reformulated to be more precise in response to some seemingly contradictory results. So, now we have HERA as a well-established set of hemodynamic hypotheses of cognitive neuroscience, which describe a certain pattern in which the human brain works when individuals engage in encoding and retrieval tasks about personally experienced events. We knew nothing of this specific pattern of brain activity before these PET and fMRI experiments were done. Thanks to the contradictory findings and the reformulation of HERA, we now have a better understanding of what the HERA model is and how it is realized by the human brain. This has involved several different groups of researchers doing experiments on HERA or related topics, interpreting these findings, and finally reformulating the HERA model. As mentioned above, the reformulated HERA has later been supported by new findings some of which came from fMRI experiments and some from experiments using newer techniques of measurement. Thus, the HERA model is a contribution of cognitive neuroscience to the general

understanding of relationships between the human brain and memory encoding and retrieval.

The above discussion of the development of the HERA model illustrates the way in which the progress of experimental knowledge goes in cognitive neuroscience. When we look at fMRI with an eye toward appreciating the kind of knowledge it can reliably provide, we can see that the more fruitful approach to fNI experiments is seeing them as tools for expanding our knowledge on relationships between hemodynamic processes and cognition rather than testing existing modularist theories. This is exactly what was done in the development of the HERA model. Furthermore, taken too far, some of these theories may impede the progress of cognitive neuroscience, because, as was discussed above, in the context of some theories HERA, a set of well-established empirical findings, does not even make sense. However, this does not necessarily mean that all theories of cognitive psychology are underdetermined by data. We have to regularly do the exercise of trying to determine what may be underdetermined and what need not be underdetermined in theories of cognitive psychology. While these theories do not strictly talk about hemodynamic processes in the brain, we can obtain knowledge about this from fNI experiments, which can then be used in partially evaluating theories. For example, any theory that excludes the HERA findings we can safely conclude to be questionable. Functional neuroimaging findings exist about cognitive processes other than memory and more findings will continue to accumulate as well-established findings. All these fNI findings can be established independently of large-scale theories but they can also be used in evaluating of existing theories of the relationships between human cognition and

the brain as well as in formulating new theories. This shows that experimental knowledge in cognitive neuroscience has “a life of its own” independently of large-scale theories of human cognition. By formulating and addressing problems of evidence and inference in fNI, the error-statistical philosophy can help this life grow.

Bibliography

Babiloni, C., Vecchio, F., Cappa, S., Pasqualetti, P., Rossi, S., Miniussi, C., Rossini, P.M. (2006). "Functional Frontoparietal Connectivity During Encoding and Retrieval Processes Follows HERA Model: A High-Resolution Study." *Brain Research Bulletin, 68,* 203 – 212.

Baier, B., Karnath, H.O., Dieterich, M., Birklein, F., Heinze, C., Muller, N.G. (2010). "Keeping Memory Clear and Stable—The Contribution of Human Basal Ganglia And Prefrontal Cortex To Working Memory." *The Journal of Neuroscience, 30,* 9788 –9792.

Bechtel, W. (2002a). "Decomposing the Mind-Brain: A Long-Term Pursuit." *Brain and Mind, 3,* 229-242.

Bechtel, W. (2002b). "Aligning multiple research techniques in cognitive neuroscience: Why is it important?" *Philosophy of Science, 69,* S48–S58.

Begley, S. (2009). "Of Voodoo and the Brain: Patterns of Neural Activity and Thoughts or Feelings Are Not as Tightly Linked as Scientists Have Claimed." *Newsweek, 153,* 52.

Bennett, C.M., Baird, A.A., Miller, M.B., & Wolford, G.L. (2010). "Neural Correlates of Interspecies Perspective Taking in the Post-Mortem Atlantic Salmon: An Argument For Proper Multiple Comparisons Correction." *Journal of Serendipitous and Unexpected Results, 1,* 1-5.

Bogen, J. (2001). “Functional Imaging Evidence: Some Epistemic Hot Spots.” In Machamer, P., McLaughlin, P., and R. Grush (eds.) *Theory and Method in the Neurosciences*. Pittsburgh, PA: University of Pittsburgh Press.

Bogen, J. (2002). “Epistemological Custard Pies from Functional Brain Imaging.” *Philosophy of Science*, *69*, S59–S71.

Bogen, J. (2010) "Theory and Observation in Science." *The Stanford Encyclopedia of Philosophy (Spring 2010 Edition)*, Edward N. Zalta (ed.), URL = <http://plato.stanford.edu/archives/spr2010/entries/science-theory-observation/>.

Buxton, R. B. (2002). *Introduction to Functional Magnetic Resonance Imaging: Principles and Techniques*. New York, NY: Cambridge University Press.

Canli, T., Sivers, H., Whitfield, S.L., Gotlib, I.H., Gabrieli, J.D.E. (2002). “Amygdala Response to Happy Faces as a Function of Extraversion.” *Science*, *296*, 2191.

Cohen, J. (1994). “The Earth is Round (p<.05).” *American Psychologist*, 49, 907-1003.

Cole, M.A. (2006). “Effects of Goal-Setting on Memory Performance in Young and Older Adults: A Functional Magnetic Resonance Imaging (fMRI) Study.” *Dissertation Abstracts International: Section B: The Sciences and Engineering*, 5134.

Coltheart, M. (2006). “What Has Functional Neuroimaging Told Us About The Mind (so Far)?” *Cortex*, *42*, 323-331.

Coltheart, M. (2010). “What Is Functional Neuroimaging For?” In Hanson, S. J. & M. Bunzl (eds.) *Foundational Issues in Human Brain Mapping*. Cambridge, MA: The MIT Press.

Damadian, R. (1971). “Tumor Detection by Nuclear Magnetic Resonance.” Science, *171, 3976*, 1151-1153.

De Haën, C. (2001). “Conception of the First Magnetic Resonance Imaging Contrast Agents: A Brief History.” *Topics in Magnetic Resonance Imaging, 12*, 221-230.

Diener, E. (2009). “Editor’s Introduction to Vul et al. (2009) and Comments.” *Perspectives on Psychological Science, 4 (3), 272-273.*

Duhem, P. (1906/1991). *The Aim And Structure of Physical Theory.* [Translated from the French by Philip P. Wiener] Princeton, NJ: Princeton University Press.

Duncan, D. E. (May, 2009). “Experimental Brain.” *Discover, May 2009,* 64 – 70, 75.

Fransson, P., Merboldt, K.D., Petersson, K.M., Ingvar, M., Frahm, J. (2002). “On the Effects of Spatial Filtering—A Comparative fMRI Study of Episodic Memory Encoding at High and Low Resolution.” *NeuroImage*, *16*, 977–984.

Gabrieli, J.D.E., Desmond, J.E., Demb, J.B., Wagner, A.D., Stone, M.V., Vaidya, C.J., & Glover, G.H. (1996). “Functional Magnetic Resonance Imaging Of Semantic Memory Processes In The Frontal Lobes.” *Psychological Science, 7*, 278-283.

Gabrieli, J.D.E., Poldrack, R.A., & Desmond, J.E. (1998). “The Role of Left Prefrontal Cortex in Language and Memory.” *Proceedings of the National Academy of Sciences*, *95*, 906 – 913.

Gazzaniga, M.S. (2000). *Cognitive Neuroscience: A Reader*. Oxford, UK: Blackwell Publishers, Ltd.

Gazzaniga, M.S. (2004). *The Cognitive Neurosciences III: Third Edition*. Cambridge, MA: The MIT Press.

Gazzaniga, M.S. (2009). *The Cognitive Neurosciences: Fourth Edition*. Cambridge, MA: The MIT Press.

Habib, R., Nyberg, L., & Tulving, E. (2003). "Hemispheric Asymmetries of Memory: The HERA Model Revisited." *Trends in Cognitive Sciences, 7,* 241 – 245.

Hacking, I. (1983). *Representing and Intervening: Introductory Topics in the Philosophy Of Natural Science*. Cambridge, UK: Cambridge University Press.

Hagen, R.L. (1997). "In Praise of the Null Hypothesis Statistical Test." *American Psychologist,* 52, 15-24.

Hagen, R.L. (1998). "A Further Look at Wrong Reasons to Abandon Statistical Testing." *American Psychologist,* 53, 801-803.

Hanson, N.R. (1958). *Patterns of Discovery*. Cambridge, UK: Cambridge University Press.

Hardcastle, V.G. & Stewart, C.M. (2002). "What Do Brain Data Really Show?" *Philosophy of Science, 69,* S72-S82.

Hashemi, R.H., Bradley, Jr., W.G., & Lisanti, C.J. (2010). *MRI: The Basics,* Third Edition. Philadelphia, PA: Lippincott Williams & Wilkins.

Henson, R. (2005). "What Can Functional Neuroimaging Tell The Experimental Psychologist." *The Quarterly Journal of Experimental Psychology, 58A,* 193-233.

Hill, L. (1896). *The Physiology and Pathology of the Cerebral Circulation: An Experimental Research*. London, UK: Churchill.

Huettel, S.A., Song, A.W., & McCarthy, G. (2004). *Functional Magnetic Resonance Imaging*. Sunderland, MA: Sinauer Associates, Inc. Publishers.

Huettel, S.A., Song, A.W., & McCarthy, G. (2008). *Functional Magnetic Resonance Imaging, 2nd Edition*. Sunderland, MA: Sinauer Associates, Inc. Publishers.

James, W. (1890). *The Principles of Psychology*. New York, NY: Henry Holt and Co.

Kelley, W.L., Miezin, F.M., McDermott, K., Buckner, R.L., Raichle, M.E., Cohen, N.J., & Petersen, S.E. (1998). "Hemispheric Specialization in Human Dorsal Frontal Cortex and Medial Temporal Lobes for Verbal and Nonverbal Memory Encoding." *Neuron, 20*, 927 – 936.

Klein, C. (2010a). "Philosophical Issues in Neuroimaging." *Philosophy Compass, 5*, 186-198.

Klein, C. (2010b). "Images Are Not the Evidence in Neuroimaging." *British Journal for the Philosophy of Science, 61*, 265–278.

Krantz, D.H. (1999). "The Null Hypothesis Testing Controversy in Psychology." *Journal of the American Statistical Association*, 44, 1372-1382.

Kuhn, T. (1962/1996). *The Structure of Scientific Revolutions* (Third Ed.). Chicago, IL: The University of Chicago Press.

LaConte, S., Anderson, J., Muley, S., Ashe, J., Frutiger, S., Rehm, K., Hansen, L.K., Yacoub, E., Hu, X., Rottenberg, D., & Strother, S. (2003). "The Evaluation of Preprocessing Choices in Single-Subject BOLD fMRI Using NPAIRS Performance Metrics." *NeuroImage, 18*, 10–27.

Lauterbur, P. C. (1973). "Image Formation By Induced Local Interactions: Examples Employing Nuclear Magnetic Resonance." *Nature, 242*, 190-191.

Lazar, N. A. (2008). *The Statistical Analysis of Functional MRI Data*. New York, NY: Springer.

Lazar, N. A. (2009). "Discussion of 'Puzzlingly High Correlations in fMRI Studies of Emotion, Personality, and Social Cognition' by Vul et al. (2009)." *Perspectives on Psychological Science, 4 (3),* 308-309.

Lee, J.H., Durand, R., Gradinaru, V., Zhang, F., Goshen, I., Kim, D.S., Fenno, L.E., Ramakrishnan, C., & Deisseroth, K. (2010). "Global and Local fMRI Signals Driven By Neurons Defined Optogenetically By Type and Wiring." *Nature, 465,* 788-792.

Lieberman, M.D., Berkman, E.T., & Wager, T.D. (2009). "Correlations in Social Neuroscience Aren't Voodoo." *Perspectives on Psychological Science, 4 (3),* 299-307.

Lindquist, M.A. & Gelman, A. (2009). "Correlations and Multiple Comparisons in Functional Imaging: A Statistical Perspective (Commentary on Vul et al., 2009)." *Perspectives on Psychological Science, 4 (3),* 310-313.

Logothetis, N. K. (2008). "What We Can Do and What We Cannot Do With fMRI." *Nature, 453,* 869- 878.

MacCorquodale, K. & Meehl, P.E. (1948). "On A Distinction Between Hypothetical Constructs and Intervening Variables." *Psychological Review, 55,* 95-107.

Machery, E. (forthcoming). "Philosophy of Psychology." In Fritz Allhoff (Ed.), *Philosophy of the Special Sciences. SUNY Press.*

Mattson, J., Simon, M. (1996). *The Pioneers of NMR and Magnetic Resonance in Medicine : The Story of MRI.* Ramat Gan, Israel: Bar-Ilan University Press.

Mayo, D. (1996). *Error and the Growth of Experimental Knowledge.* Chicago, IL: The University of Chicago Press.

Mayo, D. (2005a). "Evidence as Passing Severe Tests: Highly Probable versus Highly Probed Hypotheses." in Achinstein, P. (ed.) *Scientific Evidence: Philosophical Theories & Applications,* pp. 95 – 127. Baltimore, MD: The Johns Hopkins University Press.

Mayo, D. (2005b). "Philosophy of Statistics." in S. Sarkar and J. Pfeifer (eds.) *Philosophy of Science: An Encyclopedia*, 802-815. London, UK: Routledge.

Mayo, D. (2008). "How to Discount Double-Counting When It Counts: Some Clarifications." *British Journal of Philosophy of Science, 59*, 857-879.

Mayo, D. & Spanos, A. (2004). "Methodology In Practice: Statistical Misspecification Testing." *Philosophy of Science, 71*, 1007-1025.

Mayo, D. & Spanos, A. (2006). "Severe Testing as a Basic Concept in a Neyman-Pearson Philosophy of Induction." *British Journal of Philosophy of Science, 57*, 323-357.

Mayo, D. & Spanos, A. (2010). *Error and Inference: Recent Exchanges on Experimental Reasoning, Reliability, and the Objectivity of Science.* New York, NY: Cambridge University Press.

Mayo, D. & Spanos, A. (2011). "Error Statistics." In Prasanta S. Bandyopadhyay and Malcolm Forster (eds.) *The Handbook of Philosophy of Science, Volume7: Philosophy of Statistics*. Amsterdam, The Netherlands: Elsevier Publishers.

Mazziotta, J. C., Toga, A. W., Evans, A., Fox, P., & Lancaster, J. (The International Consortium of Brain Mapping, ICBM). (1995). "A Probabilistic Atlas of the Human Brain: Theory and Rationale for Its Development." *Neuroimage, 2*, 89-101.

Meehl, P.E. (1967). "Theory-testing in Psychology and Physics: A Methodological Paradox." *Philosophy of Science, 34*, 103-115.

Meehl, P.E. (1991). *Selected Philosophical and Methodological Papers*. C.A. Anderson & K.G. Gunderson, eds. Minneapolis, MN: University of Minnesota Press.

Mosso, A. (1881). *Ueber den Kreislauf des Blutes im Menschlichen Gehirn*. Leipzig, Germany: Von Veit.

Nichols, T.E. & Poline, J.B. (2009). "Commentary on Vul et al.'s (2009) 'Puzzlingly High Correlations in fMRI Studies of Emotion, Personality, and Social Cognition'." *Perspectives on Psychological Science, 4 (3),* 291-293.

Nyberg, L., Cabeza, R., & Tulving, E. (1996). "PET Studies of Encoding and Retrieval: The HERA Model." *Psychonomic Bulletin and Review, 3,* 135 – 148.

Ogawa, S., Lee, T.M., Kay, A.R., & Tank, D.W. (1990). "Brain Magnetic Resonance Imaging With Contrast Dependent On Blood Oxygenation." *Proceedings of the National Academy of Sciences, 87,* 9868-9872.

Okamoto, M., Wada, Y., Yamaguchi, Y., Kyutoku, Y., Clowney, L., Singh, A.K., Dan, I. (2011). "Process-specific Prefrontal Contributions to Episodic Encoding and Retrieval of Tastes: A Functional NIRS Study." *NeuroImage, 54,* 1578 – 1588.

Pauling, L. & Coryell, C.D. (1936). "The Magnetic Properties and Structure of Hemoglobin, Oxyhemoglobin and Carbonmonoxyhemoglobin." *Proceedings of the National Academy of Sciences, 22 (4),* 210-216.

Petersson, K.M., Nichols, T.E., Poline, J.B., & Holmes, A.P. (1999). "Statistical Limitations in Functional Neuroimaging I. Non-inferential Methods and Statistical Models." *Philosophical Transactions of the Royal Society of London, 354,* 1239-1260.

Provost, J.S., Petrides, M., & Monchi, O. (2010). "Dissociating The Role Of The Caudate Nucleus And Dorsolateral Prefrontal Cortex In The Monitoring Of Events Within Human Working Memory." *European Journal of Neuroscience, 32,* 873-880.

Purcell, E.M., Torrey, H.C., Pound, R.V. (1946). "Resonance Absorption By Nuclear Magnetic Moments In A Solid." *Physical Review, 69,* 37-38.

Raichle, M. E. (1998). "Behind the Scenes of Functional Brain Imaging: A Historical and Physiological Perspective." *Proceedings of the National Academy of Sciences,* 95, 765-772.

Roberts, J. A. (2002). *Instruments and Domains of Knowledge: The Case of Nuclear Magnetic Resonance Spectroscopy, 1956 – 1969.*Master's Thesis, Virginia Polytechnic Institute and State University.

Roskies, A. (2008). "Neuroimaging and Inferential Distance." *Neuroethics, 1*, 19-30.

Roskies, A. (2010). "Neuroimaging and Inferential Distance." In Hanson, S. J. & M. Bunzl (eds.) *Foundational Issues in Human Brain Mapping*. Cambridge, MA: The MIT Press.

Roy, C.S. & Sherrington, C.S. (1890). "On The Regulation of The Blood-Supply Of The Brain." *Journal of Physiology, 11*, 85-108, 158-7-158-17.

Rozeboom, W.W. (1960). "The Fallacy of the Null Hypothesis Significance Test." *Psychological Bulletin*, 57, 416-428.

Savoy, R.L. (2001). "History and Future Directions of Human Brain Mapping and Functional Neuroimaging." *Acta Psychologica, 107,* 9-42.

Shattuck, D.W., Mirza, M., Adisetiyo, V., Hojatkashani, C., Salamon, G., Narr, K.L., Poldrack, R.A., Bilder, R.M., Toga, A.W. “Construction of a 3D Probabilistic Atlas of Human Cortical Structures.” *NeuroImage*, *39*, 1064-1080.

Spanos, A. (1998). *Probability Theory and Statistical Inference: Econometric Modeling with Observational Data*. Cambridge, UK: Cambridge University Press.

Squire L. R. & E. R. Kandel. (1999). *Memory: From Mind to Molecules*. New York, NY: Scientific American Library.

Stark, C.E.L., & Squire, L.R. (2000). "Functional Magnetic Resonance Imaging (fMRI) Activity in the Hippocampal Region during Recognition Memory." *The Journal of Neuroscience, 20 (20)*, 7776-7781.

Stark, C.E.L., & Squire, L.R. (2001). “When Zero Is Not Zero: The Problem of Ambiguous Baseline Conditions In fMRI.” *Proceedings of the National Academy of Sciences*, *98*, 12760–12766.

Strother, S.C., Anderson, J., Hansen, L.K., Kjems, U., Kustra, R., Sidtis, J., Frutiger, S., Muley, S., LaConte, S., & Rottenberg, D. (2002). “The Quantitative Evaluation of Functional Neuroimaging Experiments: The NPAIRS Data Analysis Framework.” *NeuroImage*, *15*, 747–771.

Talairach, J., & Tournoux, P. (1988). *Co-planar Stereotaxic Atlas of the Human Brain*. New York, NY: Thieme Medical Publishers.

Thimm, M., Krug, A., Markov, V., Krach, S., Jansen, A., Zerres, K., Eggermann, T., Stocker, T., Shah, N.J., Nothen, M.M., Rietschel, M., & Kircher, T. (2010). “The Impact of Dystrobrevin-Binding Protein I (DTNBPI) on Neural Correlates of

Episodic Memory Encoding and Retrieval." *Human Brain Mapping*, *31*, 203 – 209.

Thulborn, K. R., Waterton, J. C., Matthews, P. M. & Radda, G. K. (1982). "Oxygen Dependence of the Transverse Relaxation Time of Water Protons in Whole Blood at High Field." *Biochimica et Biophysica Acta, 714*, 265-270.

Tulving, E., Kapur, S., Craik, F.I.M., Moscovitch, M., Houle, S. (1994). "Hemispheric Encoding/Retrieval Asymmetry in Episodic Memory: Positron Emission Tomography Findings." *Proceedings of the National Academy of Sciences*, *91*, 2016 – 2020.

Uttal, W. (2001). *The New Phrenology: The Limits of Localizing Cognitive Processes in the Brain*. Cambridge, MA: MIT Press.

Uttal, W. (2002a). "Functional Brain Mapping – What Is It Good For? Plenty, but not Everything! (A Reply to Malcolm J. Avison)." *Brain and Mind*, 3, 375-379.

Uttal, W. (2002b). "Précis of The New Phrenology: The Limits of Localizing Cognitive Processes in the Brain." *Brain and Mind*, 3, 221-228.

Van Orden, G.C., & K.R. Paap. (1997). "Functional Neuroimages Fail To Discover Pieces of Mind In Parts of The Brain." *Philosophy of Science, 64*, S85–S94.

Vul, E., Harris, C., Winkielman, P., & Pashler, H. (2009). "Puzzlingly High Correlations in fMRI Studies of Emotion, Personality, and Social Cognition" *Perspectives on Psychological Science, 4 (3)*, 274-290.

Watson, J. (1913). "Psychology as the Behaviorist Views It." *Psychological Review*, *20*, 158-177.

Watson, J. (1928). *The Ways of Behaviorism*. New York, NY: Harper & Brothers Pub.

Wilkinson, L. and the Task Force on Statistical Inference. (1999). "Statistical Methods in Psychology Journals: Guidelines and Explanations." *American Psychologist*, 54, 594-604.

Yarkoni, T. (2009). "Big Correlation in Little Studies: Inflated fMRI Correlations Reflect Low Statistical Power—Commentary on Vul et al. (2009)" *Perspectives on Psychological Science, 4 (3),* 294-298.