Functional Imaging in Behavioral Neurology and Cognitive Neuropsychology

Geoffrey Karl Aguirre

From: T. E. Feinberg & M. J. Farah (Eds.), Behavioral Neurology and Cognitive Neuropsychology . (in press). New York: McGraw Hill.

Introduction

What can we hope to learn of the physical operation of the central nervous system and the mental processes that result? The fields of behavioral neurology and cognitive neuropsychology survey this relationship, and have at their disposal powerful intellectual and methodological tools. While the terrain of possible models of brain and behavior interaction seems boundless, particular tests of these models may exist beyond the reach of particular methods.

These methods fall into two broad categories, and have been around for several centuries. The first category includes manipulations of the neural substrate itself. Such an intervention might inactivate a brain area, perhaps through a lesion, with Paul Broca's 1861 observation of the link between language and left frontal lobe damage providing a prototypical example. The effects of stimulation of brain areas can also be studied, as Harvey Cushing did with the human sensory cortex in the early 20th century. In contrast, observation techniques relate a measure of neural function to behavior. Hans Berger's work in the 1920s on the human electroencephalographic response is a good starting point.

Impressive refinements and additions to both of these categories have taken place over the last century. For example, beyond the static lesions of "nature's accidents" that have been the mainstay of cognitive neuropsychology for many years, it is now possible to temporarily and reversibly inactivate areas of human cortex using transcranial magnetic stimulation (see chapter X for additional details). The realm of "observational" methodology has also grown dramatically in the last few decades, with the development of functional neuroimaging.

In this chapter we concern ourselves with the theoretical and practical properties of functional neuroimaging techniques in general, and with BOLD fMRI in particular. I'll begin with a brief consideration of the nuts-and-bolts physics and physiology that underlie common imaging methods. Next, we'll consider several aspects of the inferential basis of neuroimaging. Our deliberations here will include a "two systems" model of neuroimaging inference, and three

general types of hypotheses that can be tested using these methods. We'll also explore the relationship between neuroimaging and lesion studies and describe different methods for behaviorally isolating cognitive processes of interest within a neuroimaging experiment. Finally, I'll discuss some idiosyncratic properties of BOLD fMRI as they relate to different categories of temporal organization of experiments (e.g., blocked, event-related, etc.). Except for a few glancing references, the subject of the statistical analysis of neuroimaging data in general and BOLD fMRI in particular will be avoided.

Properties of functional neuroimaging data

In general, functional neuroimaging can be defined as the class of techniques that provide volumetric, spatially localized measures of neural activity from across the brain and across time; in essence, a three-dimensional movie of the active brain. Importantly, functional imaging data have a particular order in time which cannot be capriciously changed without fundamentally altering the nature of the original data. Virtually all neuroimaging experiments vary an experimental condition over time and evaluate the relationship between the experimental manipulation and the observed time series.

Relatively non-invasive measurements of blood flow in the human brain was first accomplished in 1963 by Glass and Harper by measuring the decay of inhaled, radioactive xenon gas. This method could only provide global measurements of blood flow within the head, so could not be used to generate images. This was to change soon after the introduction of computed axial tomography (CT) in the 1970s. Developed for use with x-ray images, CT methods allowed the reconstruction of a volumetric image of the body by passing x-rays through the subject from multiple directions. These ideas were soon applied to measurements of metabolic function in the human brain, with the twist that instead of directing radiation energy through the body, the source of radiation was located within the subject. In positron emission tomography (PET) scanning, the subject is injected with a radioisotope that, as it decays, produces positrons. These immediately annihilate with electrons, producing two photons that travel outward in opposite directions. An array of sensors located around the subject's head uses "coincidence detection" to determine the source of the radioactive decay in space.

Using radio-labeled water, PET initially saw application as a method of measuring local cerebral tissue perfusion. In the 1980s, functional changes in blood flow became the object of

study, by comparing the distribution of cerebral perfusion during two different cognitive states. In the following decades, PET techniques were used extensively in the service of cognitive neuropsychology research. While PET provides for a spatial resolution on the order of a few millimeters, temporal resolution is limited by the half-life of the radioisotope used. Practically, PET images can only be obtained every few minutes, limiting the ability of the method to dynamically track changes in neural activity related to cognitive processes. It is this limitation in temporal resolution, coupled with the invasive and expensive need for radioisotope injection, that led ultimately to fMRI methods supplanting PET as the primary tool of cognitive neuroscience investigation.

The use of magnetic resonance imaging for the assay of neural function was initiated by Belliveau and his colleagues in 1991, who used an injected contrast agent (gadolinium) to obtain perfusion MRI images of the occipital cortex during visual stimulation. The widespread application of functional MRI awaited the development of a non-invasive, endogenous tracer method, subsequently introduced by Ogawa and colleagues in 1993. Fortuitously, hemoglobin, the primary oxygen carrying molecule in the blood, has different magnetic properties when bound and unbound to oxygen, and this serves as the agent of contrast in blood oxygen level dependent (BOLD) fMRI. Local changes in neural activity give rise to a chain of physiologic and imaging events, many of the details of which are still under study. In brief, increases in neural activity produce local increases in blood flow (Leniger-Follert et al., 1979), which in turn engender a delayed decrease in local deoxyhemoglobin concentration (Malonek et al., 1996). This is sometimes referred to as a paradoxical change, as increased metabolic activity leads to a decrease in deoxyhemoglobin. Because deoxyhemoglobin has stronger magnetic properties than oxyhemoglobin, a decrease in the deoxyhemoglobin concentration results in a decreased pertubation of the local magnetic field (referred to as a *susceptibility* gradient). This increases the T2* weighted fMRI signal (Jezzard et al., 1996) from the area, which serves as the dependent data for fMRI. Several excellent reviews of the physics and physiology underlying BOLD fMRI are available for the interested reader (e.g., Moonen and Bandettini, 1999).

Unlike PET where the signal measured can be expressed as a physical quantity (e.g., cc of blood / 100 g of tissue / minute), the BOLD fMRI signal has no simple, absolute interpretation. This is because the particular signal value obtained is not *exactly* a measure of

deoxyhemoglobin concentration, but is instead a measure which is *weighted* by this concentration (i.e. is T2* weighted) and is also influenced by a number of other factors that can vary from voxel to voxel, scan to scan, and subject to subject. As a result, experiments conducted with BOLD fMRI generally test for differences in the magnitude of the signal between different conditions within a scan. One could not, for example, directly contrast the absolute level of the BOLD fMRI signal obtained within the temporal lobe of schizophrenic patients with that from controls with much hope of obtaining a reliable or unbiased statistical test. Notably, recent developments in perfusion imaging offer the ability to obtain an fMRI signal which can be interpreted in concrete physical units (Detre and Alsop, 1999; Aguirre et al, 2002).

The spatial and temporal resolution of BOLD fMRI is limited by the neuro-vascular coupling that is the source of contrast. While MRI images can readily be obtained every 100 msecs, and with spatial resolution in the tenths of a millimeter, this fine resolution has little practical advantage. Changes in neural activity give rise to a change in BOLD fMRI signal that evolves over seconds (described in detail below). As a result, BOLD fMRI images are seldom acquired more frequently than once a second. Additionally, a point of neural activity engenders a change in BOLD signal that spreads over several millimeters, thus BOLD images are typically composed of voxels (the smallest volume "pixel" of which the image is composed) no smaller than 1 mm on a side.

The "two-systems" model of functional neuroimaging experiments

Engineers frequently find it convenient to discuss an object of study as a *system*. Simply defined, a system is something that takes input and provides output. One can imagine many different examples of systems, such as a car, where the pressure upon the gas pedal is the input, and the speed of the car is the output. Certain types of systems are particularly amenable to study and characterization.

What is the system under study in functional neuroimaging experiments of cognitive neuropsychology hypotheses? A profitable way of answering this question is with a *two systems* model, in which the domain of mental operations and cognitive processes is held separate from vascular physiology and imaging physics (Figure 1).

The first system is that of cognition, in which the inputs are the instructions, stimuli and tasks presented to the subject by the experimenter, and the output is the pattern of neural activity

evoked within the brain. The second system is the domain of physiology and physics, and mediates the transformation of neural activity inputs into blood flow responses and imaging signal. In most cognitive neuroscience experiments, the hypotheses of interest concern the patterns of neural activity that are evoked by cognitive processes, and therefore chiefly concern the first of these two systems. Of course, many scientists have great interest in the exact physical mechanisms that mediate the neuro-vascular response, and the second system is therefore a frequent target of experiments that do not strictly fall within the realm of cognitive neuroscience.

If it were the case that the properties of both systems were unknown, then it would be a daunting (if not impossible) task to study cognition using functional neuroimaging. This is because one would not be able to assign a given change in imaging signal to cognition or neuro-vascular coupling. Fortunately, the properties of the second system are lawful and well described, even if the exact mechanisms of the transformation are still not well understood. Therefore, one is able state what changes in neural activity are implied by a given pattern of imaging signal. After deriving the implied pattern of neural activity from the observed signal, inferences can be drawn regarding the relationship between cognition and neural activity. This is the process that we effortlessly (and frequently unconsciously) engage when we look at a neuroimaging statistical map.

It is essential to keep the distinction between these two systems clear when considering functional neuroimaging experiments and their interpretation, as there are properties that can be ascribed to one system that are clearly not appropriate for the other. For example, and as will be discussed in greater depth below, it has been demonstrated that the system that transforms neural activity into BOLD fMRI signal is nearly *linear*. For example, twice the neural input leads to twice the BOLD fMRI signal output. This is obviously not a property that can be readily assumed to be true of the cognition system—presenting twice as many words to be remembered is not *a priori* assumed to produce twice as much neural activity (although this is a property of the system which might be tested).

What aspect of the first system is studied in a neuroimaging experiment? Typically, these experiments involve a subject in the scanner performing a task of the design of the experimenter. (Although there are certainly exceptions: consider studies of seizure activity, REM sleep). As depicted, it is the stimuli and instructions from the task that constitute the inputs to the first

system. Ultimately, the purpose of the experiment is to state some relationship between the neural activity observed and the behavior in which the subject engaged. Therefore, the researcher often seeks to control not only the *task* that the subject performs, but the *mental states* that the subject enters. The internal mental state(s) of the subject are typically referred to as the cognitive processes. Importantly, a cognitive process is distinct from a task, in that multiple tasks might be thought to be able to evoke a single cognitive process (see table 1).

Note that simply stating that something is a cognitive process does not make it so! The notion of a cognitive process has a fairly rigorous instantiation within psychology, and the demonstration of the existence of a particular cognitive process is often the target of much behavioral research. For example, Sternberg's additive factors method is a logical system used to identify task manipulations that can demonstrate the existence of independent cognitive processes (Sternberg 1969).

Basic types of neuroimaging inference

Although there are many neuroimaging methods and a seemingly limitless number of applications, the basic type of question being asked usually fits within one of a few categories. Each type of approach makes slightly different assumptions and permits different inferences about the relationship between the brain and behavior. Worth noting now, and as will be developed, an experiment that tries to fit into more than one category at once is likely on shaky logical ground.

By far the most common application of neuroimaging methodology is to *localization* questions, which ask: what are the neural correlates of a given cognitive process? Generally, the subject is presented with a task designed to selectively evoke a particular cognitive state of interest, and the neuroimaging method identifies if and where bulk changes in neural activity accompany that cognitive process. The key assumption for this type of design is that a given cognitive process exists and that the task isolates only that cognitive process. Various techniques are used (such as cognitive subtraction or parametric manipulation, discussed below) to isolate the mental operation of interest from the other processes that invariably are present (e.g., button pushing, preparing responses, etc.).

It is important to understand what these type of experiments *cannot* conclude. Activity evoked by a particular cognitive process cannot be taken as evidence that the activated cortical

region is necessary for the cognitive process (in the sense that a lesion of the area would impair the subject's ability to perform the task). If the assumptions of the localization framework are perfectly met, the strongest inference that can be made is that the region is *activated* by the cognitive process. Demonstration of necessity requires a lesion study (or some other method of inactivating neural structures such as transcranial magnetic stimulation; see below).

In contrast, *implementation* studies ask about the computational mechanisms of a cognitive process within a cortical region. This type of study begins with the assumption that a cortical region is involved in a particular cognitive process. The purpose of the study is then to determine the neuro-computational parameters that mediate the area's participation in that process. For example, one might know that area MT is involved in the cognitive process of motion detection. However, what is the relationship between speed of motion and the MT response? Does motion directed toward the viewer, as opposed to across the visual field, change the magnitude of neural activity? As another example, consider a region of prefrontal cortex assumed to be involved in the cognitive process of working memory. Does this region change its bulk level of neural activity with increasing memory load (i.e., remember four items instead of two)?

Finally, an *evocation* design turns the familiar direction of neuroimaging inference on its head and asks: what cognitive process does a given task evoke? Also termed "reverse inference", this framework is used to make inferences about cognition, as opposed to neural activity; i.e., the behavior of the subject is the unknown variable. One begins by assuming that a particular cortical region is activated by a *single cognitive process*. This mapping must be unique, in that one and only one cognitive process is capable of activating a particular region. The subject performs a task which may or may not evoke the cognitive process of interest. The fMRI data are then examined to determine if increased neural activity was present within the specified region during the task and, if so, the conclusion is drawn that the subject recruited the cognitive function. In other words, the evocation paradigm may be used to test hypotheses regarding the engagement of cognitive processes during a behavioral state in which the cognitive processes need not be under experimental control.

Suppose, for example, we assume that neural activity in area MT indicates the presence of the cognitive process of motion perception. By examining the neural activity within this area,

it becomes possible to learn how unrelated distractors affect motion perception (Rees 1997). In another example, we might assume that activity in the "fusiform face area" indicates the cognitive process of face perception. We can then monitor the responses of this area during a binocular-rivalry paradigm that pits face stimuli against house stimuli to learn about the time course of perceptual switching (Tong 1998).

What provides the evidence that a particular region is uniquely activated by a specific cognitive process? Logically, only an exhaustive neuroimaging examination of every possible cognitive process, under every possible circumstance, could provide the necessary evidence. This is obviously practically impossible, so a series of neuroimaging experiments that demonstrate activation of a particular region during a given cognitive process and no other usually suffices to support the assumption (a logical inference termed *enumerative induction*).

It is worth noting that a common logical error in neuroimaging studies is to try and conduct both evocation *and* localization inferences at the same time. Often, the discussion section of a paper will identify activity in one cortical area as the consequence of a cognitive process the experimenter intended to manipulate in the task, and then in the next paragraph suggest that activity in some other location (e.g., the frontal lobe) is the result of some other behavior in which the subject engaged (e.g., working memory). This is an error because the assumptions of each type of inferential framework contradict the other. The localization framework assumes that only a single cognitive process is being manipulated, while the evocation framework assumes that multiple, unspecified mental states are in play.

These three categories of neuroimaging inference should be taken as guidelines and do not exhaust the realm of possible designs. For example, studies of effective connectivity (Buchel and Friston, 2000) examine the relationship of neural activity in different areas of the brain, and are often used to support inferences about implementation, although other applications are certainly possible as well. The primary use of this tripartite division is to help organize one's thinking about the assumptions that underlie a particular experiment.

The relationship between lesion and neuroimaging studies

Although perhaps counterintuitive, lesion and neuroimaging studies provide for very different and non-overlapping inferences about the relationships between the brain and behavior. One way to think about these differences is in terms of the logical construct of *necessity*. A

common neuropsychological hypothesis states that a particular brain region is necessary for the performance of a particular mental operation. If the region in question is removed (perhaps through a lesion) and the cognitive process is impaired, the hypothesis has been affirmed. While this seems straightforward, several inferential complications can ensue, and the interested reader is directed to chapter X for a discussion of the lesion method. One wrinkle that will be discussed here is the possibility that more than one region is capable of supporting the process of interest (perhaps working in parallel, or one serving as a "backup" for the other). In this case, the region still plays an interesting role in the cognitive process, although it is not strictly *necessary* in that a lesion of only that region will not impair the cognitive process. Therefore, an alternate relationship between a neural substrate and a behavior that might be sought is *involvement* (I am indebted to my colleague Eric Zarahn for this particular intellectual construct and to his general contribution to the ideas in this section). A region is *involved* in a cognitive process if it is necessary under some circumstance (the circumstance being that all the other potential "backup" regions have also been damaged).

Does finding activation of a cortical region in a functional neuroimaging study imply that the region is necessary (or even just involved) in the cognitive process? In short, the answer is no; not only for functional neuroimaging but for all "observation" methods mentioned at the outset (EEG, depth electrode recording, etc). The primary cause of this state of affairs is the observational, correlative nature of neuroimaging. Although we make inferences regarding cognitive processes, these processes are not themselves directly subject to experimental manipulations. Instead, the investigator controls the presentation of stimuli and instructions, with the hope that these circumstances will provoke the subject to enter a certain cognitive state and *no other*. Careful consideration reveals how this assumption might fail. Although cooperative, the subject may unwittingly engage in confounding cognitive processes in addition to that intended by the experimenter, or alternatively, may fail to differentially engage the process. For example, a subject might constantly engage in the process of declarative memory formation, even during periods of time when he is "supposed" to be performing some other, control behavior. It is therefore not possible to know if observed changes in neural activity in a brain region are the result of the evocation of the cognitive process of interest or an unintended, confounding process. Negative results (even in the face of arbitrarily high statistical power) are also not conclusive, not only because of the failure of perfect control of evocation of cognitive

processes, but because of the possibility that the neuroimaging method employed is not sensitive to the critical change in metabolic activity (e.g., pattern of neuronal firing as opposed to bulk, integrated synaptic activity).

What about the converse inference? If a region is involved in (or necessary for) a cognitive process, may we deduce that it would be activated by that cognitive process in a neuroimaging experiment? Again, the answer is no. Consider the situation in which two cortical regions are both involved in the same cognitive process. One is the "primary" region, and the other is the "back-up" region. As long as the primary region is functioning, the back-up region is quiescent. Thus, a neuroimaging study might fail to demonstrate activation of the back-up region, even though it is involved in the cognitive process. Interestingly, one might study a patient with a lesion of the primary region, and thereby demonstrate activation of the back-up region by the cognitive process of interest under that circumstance (see below).

What about the case in which a region is actually *necessary* for a cognitive process? Can we now expect that the region will be activated by the cognitive process? The answer is yes, but with a number of caveats. For example, we must be able to assume that the necessity of the region is not some side-effect of the lesion method itself, for example damaging axons that simply passed through the area (so-called "fibers of passage" ; Jarrard 1993), or impacting the metabolic activity of remote areas (causing, for example, diaschisis; Feeney and Baron 1986). Also, we must assume that our neuroimaging method is sensitive to all possible changes in neural activity that might be induced by the cognitive process (not just bulk neural activity, but synchronicity of firing). If not, then it is possible that the necessary cortical region undergoes a change in state associated with the cognitive process that our technique is unable to observe.

The upshot is that caution should be exercised in applying neuroimaging results to the clinical interpretation of the necessity of a cortical region for behavior. For example, there has been interest in using BOLD fMRI to replace the "Wada" or intracarotid amobarbital test (Binder et al, 1996). Performed to guide surgical resection of epileptic foci, each internal carotid artery is in turn catheterized and instilled with anesthetic to determine which hemisphere is dominant for language. The hope is that BOLD fMRI can be used to determine which hemisphere responds to language tasks and replace this invasive procedure. While results so far have indicated a good correlation between the two methods, there is no logical requirement that this be the case. In fact,

counterexamples have been presented in which neuroimaging has demonstrated activation in frontal cortical areas not subsequently found to be necessary for language function, and vice-a-versa (Jayakar et al., 2002).

The comments so far have addressed hypotheses in normal, "control" subjects. What types of inferences are possible with functional neuroimaging studies of brain damaged patients? We have already mentioned one case in which a neuroimaging study of a patient population can answer an otherwise intractable question: studies of a patient with the preserved ability to perform a cognitive operation despite damage to a region thought to be involved in the task can help identify other, candidate, involved regions.

The key to this application is that the patient has *preserved functioning* of the cognitive operation of interest. In most cases, however, patients are of interest to cognitive neuropsychologists because of their impairment in a particular cognitive process. It is this very property that renders neuroimaging studies of patient populations an inferential tangle. If a patient is not performing a cognitive operation normally, what can be learned by placing them in the scanner and asking them to attempt to perform the task? Clearly, experimental designs of the localization and implementation variety will be of questionable value, as those designs begin with the assumption that the cognitive process of interest is under the control of the experimenter. This will certainly not be the case as the patient struggles to engage alternate strategies to solve tasks that they are unable to perform. It may be possible to conduct evocation experiments, in which the behavior of the patient is the subject of study, as opposed to the particular neural substrate. In this case, neural activity during altered performance of some task is used to test hypotheses about the particular compensatory cognitive processes that the patient uses. The interested reader is referred to (Price and Firston, 1999) for additional discussion of this topic.

Manipulation of the cognitive process

As was discussed in the setting of inferential framework, many neuroimaging experiments depend upon the isolation of a particular cognitive process for study. Specifically, the fundamental assumption of the localization category of neuroimaging inference is that the cognitive process of interest can be isolated from other mental operations, so that the neural correlates of that solitary process can be observed. Here we consider several broad classes of

manipulations designed to do this. Note that any of these techniques for evoking a particular cognitive process can be coupled with different temporal structures of designs, described in the next section.

Cognitive subtraction is the prototypical neuroimaging method for putative isolation of cognitive processes, the logic of which derives from similar arguments made in the study of reaction times (e,g. Donders). One condition of the experiment is designed to engage a particular cognitive process, such as face perception, episodic encoding, or semantic recall. This "experimental" condition is contrasted with a "control" condition that is designed to evoke all of the cognitive processes present in the experimental period except for the cognitive process of interest. Under the assumptions of "cognitive subtraction" (Posner et al. 1988), differences in neural activity between the two conditions can be attributed to the cognitive process of interest.

Cognitive subtraction assumes, as do the other manipulations described below, that the particular cognitive process of interest can be evoked uniquely. This is a fundamental inferential weakness of many cognitive neuroimaging studies. As has been discussed, although we make inferences regarding cognitive processes, these processes are not themselves directly subject to experimental manipulations. Even the cooperative subject might engage other, confounding, mental operations unintentionally, rendering this assumption invalid.

Cognitive subtraction (in neuroimaging) relies upon two additional assumptions: "pure insertion" and linearity. Pure insertion is the idea that a cognitive process can be *added* to a pre-existing set of cognitive processes without affecting them. This assumption is difficult to prove because one would need an independent measure of the preexisting processes in the absence and presence of the new process. If pure insertion fails as an assumption, a difference in neuroimaging signal between the two conditions might be observed not because of the simple addition of the cognitive process of interest, but because of an interaction between the added component and preexisting components. For example, the act of pressing a button to signal a semantic judgment may be different from pressing a button in response to a visual cue. Effects upon the imaging signal that result from this difference would be erroneously attributed to semantic judgment *per se*.

A second assumption of cognitive subtraction is that the transformation of neural activity into fMRI signal is linear. While the BOLD fMRI system has been shown to exhibit behavior

close to that of a linear system, there is some evidence for systematic departures (Boynton et al. 1996). Failures of linearity can cause adjacent neural events to produce more or less signal than those events would in isolation, rendering subtraction approaches invalid. In fact, failures of cognitive subtraction along these lines have been empirically demonstrated for working memory experiments (Zarahn et al. 1997).

Several other cognitive process manipulations have as their goal a reduction in the reliance upon the assumption of pure insertion. *Factorial* experiments (Friston et al. 1996) are designed to examine the interactions of two different, candidate cognitive processes. The scheme of the design involves (in the simplest case) four conditions, during which two different processes are evoked individually and then jointly. The proposed advantage of the design is that interactions between the two processes can be examined. The presence of an interaction is indicated if the difference in imaging signal between the presence and absence of cognitive process "A" is itself different when cognitive process "B" is present or absent. While factorial designs do provide a compelling method for gaining greater insight into the neural implementation of cognitive processes, it is a mistake to claim that such designs obviate the need for the pure insertion assumption. Interpretation of the design requires the assumption that the two cognitive processes have, indeed, been isolated. The logic by which this isolation is to occur is the same as that outlined for cognitive subtraction above. That is, process "A" and process "B" must be purely inserted into the other cognitive components that allow the experiment to evoke these processes.

The *cognitive conjunction* design (Price and Friston 1997) has also been proposed to reduce reliance upon the assumption of pure insertion. The logic of the approach is that, if one wishes to discount the possibility of an interaction (i.e., a failure of pure insertion) between the cognitive component to be added and the set of preexisting processes, one should repeat the experiment with a different set of preexisting processes and replicate the result. A rigorous implementation of this notion involves conducting a series of categorical subtraction experiments that all aim to isolate the same cognitive process. The novel twist is that the subtractions need not be complete; that is, the experimental and control conditions can differ in several cognitive processes in addition to the one of interest. The imaging data are then analyzed to identify areas that have a significant, consistent response to the putatively isolated process (i.e., a significant

main effect across subtractions in the absence of any significant interactions). Again, while this design reduces the plausibility of some failures of cognitive subtraction, it does not eliminate the possibility. In particular, some cognitive processes, by their very nature, require the evocation of an antecedent process. For example, can working memory be meaningfully present if not preceded by the presentation of a stimulus to be remembered? If not, then any cognitive conjunction design that attempts to demonstrate the presence of neural activity during a delay period will be susceptible to erroneous results due to interactions between the task manipulation and preexisting task components.

Finally, *parametric* designs offer an attractive alternative to cognitive subtraction approaches. In a parametric design, the experimenter presents a range of different levels of some parameter, and seeks to identify relationships (linear or otherwise) between imaging signal and the values that the parameter assumes. This can be done to identify the neural correlates of straightforward changes in stimulus properties or manipulations of a cognitive process. Unlike cognitive subtraction methods, parametric designs do not rely as heavily upon the assumption of pure insertion, as the cognitive process is present during all conditions.

Properties of the BOLD fMRI system that impact experimental design

So far I have described properties of all neuroimaging techniques, and considered the inferential consequences of different ways of influencing a subject's mental state. We now turn our attention to the idiosyncratic properties of one particular neuroimaging method: BOLD fMRI. I will focus here on two key properties of BOLD fMRI data that fundamentally impact the design of BOLD fMRI experiments: the hemodynamic response function and the presence of low-frequency noise.

As was mentioned earlier, the cascade of neuro-vascular events that ensue following changes in neural activity and produce changes in BOLD fMRI signal are still under investigation. Fortunately, the BOLD fMRI system has properties of a *linear system*, allowing us to ignore for the most part the messy details of physics and physiology. Like any other system, a linear system takes input and provides output (in this case, neural activity in, BOLD fMRI signal out). Importantly, a linear system can be completely characterized by a property called the impulse response function (IRF), which is the output of the system to an infinitely brief, infinitely intense input. In the context of BOLD fMRI, the hemodynamic response function

(HRF) is taken as an estimate of the IRF of the BOLD fMRI system, and is the change in BOLD fMRI signal that results from a brief (< 1 second) period of neural activity. Knowledge of the IRF can be used to predict the output of the system to any arbitrary pattern of input by a mathematical process called convolution. Therefore, knowledge of the shape of the HRF allows one to predict the BOLD fMRI signal that will result from any pattern of neural activity.

The HRF itself can be empirically measured from human subjects by studying the BOLD fMRI signal that is evoked by experimentally induced, brief periods of neural activity in known cortical areas (e.g., neural activity in the primary motor cortex in response to a button press). The shape of the HRF reflects its vascular origins (see figure 2), and rises and falls smoothly over a period of about 16 seconds. While the shape of the HRF varies significantly across subjects, it is very consistent within a subject, even across days to months (Aguirre et al, 1998). There is some evidence that the shape of the HRF varies from one region of the brain to another (perhaps from variations in neuro-vascular coupling), but this is a difficult notion to test as it is necessary to create evoked patterns of neural activity in disparate areas of the brain that can be guaranteed to be very similar.

The temporal dynamics of neural activity are quite rapid, on the order of milliseconds, but changes in blood flow take place over the course of seconds. One consequence of this, as demonstrated by the smooth shape of the HRF, is that rapid changes in neural activity are not well represented in the BOLD fMRI signal. The "temporal blurring" induced by the HRF leads to many of limitations placed on the types of experiments that can be conducted using BOLD fMRI. Specifically, the smooth shape of the HRF makes it difficult to discriminate closely spaced neural events. Despite this, it is still possible to detect: 1) brief periods of neural activity, 2) differences between neural events in a fixed order, spaced as closely as four seconds apart, 3) differences between neural events, *randomly* ordered, closely spaced (e.g., every second or less), and 4) neural onset asynchronies on the order of 100 msecs. The reason that these seemingly paradoxical experimental designs can work is that some patterns of events that occur rapidly or switch rapidly create a low-frequency "envelope"; a larger structure of pattern of alternation that can pass through the hemodynamic response function. In the next section I discuss several types of temporal structures for BOLD fMRI experiments, and consider how the shape of the HRF dictates the properties of these designs.

Another important property of BOLD fMRI data is that greater power is present at some temporal frequencies as compared to others under the null hypothesis (i.e., data collected without any experimental intervention). The power spectrum (a frequency representation) of data composed of independent observations (i.e., white noise), should be "flat", with equal power at all frequencies. When calculated for BOLD fMRI, the average power spectrum is found to contain ever increasing power at ever lower frequencies (see figure 3), often termed a 1/frequency distribution. This pattern of noise can also be called "pink", named for the color of light that would result if the corresponding amounts of red, orange, yellow, etc. of the visible light frequency spectrum were combined. The presence of noise of this type within BOLD fMRI data has two primary consequences. First, traditional parametric and non-parametric statistical tests are invalid for the analysis of BOLD fMRI data, which is why much of the analysis of BOLD fMRI data is conducted using Keith Worsley and Karl Friston's "modified" general linear model (1995), as instantiated in SPM and other statistical packages. The second impact is upon experimental design. Because of the greater noise at lower frequencies, slow changes in neural activity are more difficult to distinguish from noise.

The astute reader might note that the consequences of the shape of the hemodynamic response function and the noise properties of BOLD fMRI are at odds. Specifically, the shape of the HRF would tend to favor experimental designs that induce slow changes in neural activity, while the presence of low-frequency noise would argue for experimental designs that produce more rapid alterations in neural activity. As it happens, knowledge of the shape of the HRF and the distribution of the noise is sufficient to provide a principled answer as to how best balance these two conflicting forces.

It is worth noting that other neuroimaging methods have different data characteristics, with different consequences for experimental design. For example, perfusion fMRI is a relatively new approach that provides a non-invasive, quantifiable measure of local cerebral tissue perfusion (Detre and Alsop, 1999). Perfusion data do not suffer from the elevated, low-frequency noise present in BOLD and as a result, perfusion fMRI can be used to detect extremely long time-scale changes in neural activity (over minutes to hours to days) that would simply be indistinguishable from noise using BOLD fMRI (Aguirre et al., 2002).

Different temporal structures of BOLD fMRI experiments

As BOLD fMRI experiments by necessity include multiple task conditions (prototypically, an "experimental" and "control" period), several ways of ordering the presentation of these conditions exist. Different terms are used to describe the pattern of alternation between experimental conditions over time, and include such familiar labels as "blocked" or "event related". While these are often perceived as rather concrete categories, the distinction between blocked, event-related, and other sorts of designs is actually fairly artificial. These may be better considered as extremes along a continuum of arrangements of stimulus order. Consider every period of time during an experiment as a particular experimental condition. This includes the "inter-trial-interval" or "baseline" periods between stimulus presentations. In this setting, blocked and event-related designs are viewed simply as different ways of arranging periods of "rest" (or no stimulus) with respect to other sorts of conditions. (For a more complete exploration of these concepts, see Friston, 1999).

The prototypical fMRI experimental is a blocked approach in which two conditions alternate over the course of a scan. For most hypotheses of interest, these periods of time will not be utterly homogeneous but will consist of several trials of some kind presented together. For example, a given block might present a series of faces to be passively perceived, or a sequence of words to be remembered, or a series of pictures to which the subject must make a living/non-living judgment and press a button to indicate his response. Blocked designs have superior statistical power compared to all other experimental designs. This is because the fundamental frequency of the boxcar can be positioned at an optimal location with respect to the filtering properties of the hemodynamic response function and the low frequency noise.

Event-related designs model signal changes associated with individual trials, as opposed to blocks of trials. This makes it possible to ascribe changes in signal to particular events, allowing one to randomize stimuli, assess relationships between behavior and neural responses, and engage in retrospective assignment of trials. Conceptually, the simplest type of event-related design to consider is one which uses only a single stimulus type, and uses sufficient temporal spacing of trials to permit the complete rise and fall of the hemodynamic response to each trial; a briefly presented picture of a face once every sixteen seconds for example. Importantly, while this prototypical experiment has only one stimulus, it has *two* experimental conditions (the

stimulus and the inter-trial-interval). If one is willing to abandon the fixed ordering and spacing of these conditions, more complex designs become possible. For example, randomly ordered picture presentations and rest periods could be presented as rapidly as once a second. The ability to present rapid alternations between conditions initially seems counterintuitive, given the temporal smoothing effects of the hemodynamic response function. While BOLD fMRI is insensitive to the particular high-frequency alternation between one trial and the next, it is still sensitive to the low-frequency "envelope" of the design. In effect, with closely spaced, randomly ordered trials, one is detecting the low frequency consequences of the random assortment of trial types.

The discussion thus far regarding event-related designs has assumed an ability to randomize perfectly the order of presentation of different event types. There are certain types of behavioral paradigms, however, that do not permit a random ordering of the events. For example, the delay period of a working memory experiment always follows the presentation of a stimulus to be remembered. In this case, the different events of the trial cannot be placed arbitrarily close together without risking the possibility of false positive results that accrue from the hemodynamic response to one trial event (e.g., the stimulus presentation) being interpreted as resulting from neural activity in response to another event (e.g., the delay period). It turns out that, given the shape typically observed for hemodynamic responses, events within a trial as close together as four seconds can be reliably discriminated (Zarahn et al. 1997). Thus, event-related designs can be used to examine directly, for example, the hypothesis that certain cortical areas increase their activity during the delay period of a working memory paradigm without requiring the problematic assumptions traditionally employed in blocked, subtractive designs.

As a final example of event-related design, consider an experiment that aims to identify a neural onset asynchrony. As noted above, the hemodynamic response observed for a subject during a scanning session is highly reliable in its shape, and is relatively smooth (i.e., is not composed of high-frequency components). As a result, it might be possible to use fMRI to detect small neural onset asynchronies. Such a design might present one of two different behavioral trials (in a random order) every 16 seconds. Because of the reliability of the hemodynamic transformation, differences in the mean time-to-peak of the responses to the two different types of stimuli could be identified and ascribed to an asynchrony in onset of neural activity (Menon et

al., 1998). Such a design might allow one, for example, to test the hypothesis that a cortical area that responds to pictures of faces responds with a slightly longer latency (on the order of 100 msec) to pictures of inverted faces.

Conclusion

There is an enormous variety of experimental designs that may be used to ask cognitive neuropsychology questions with neuroimaging methods. I have provided here frameworks for organizing these approaches into inferential categories, methods of manipulating evoked cognitive processes, and ways of arranging different experimental conditions in time. Hopefully, the general principles enumerated here will be of use not only for BOLD fMRI, the preeminent neuroimaging method of today, but for whatever leap in neuroimaging methodology awaits us tomorrow.

References

Aguirre GK, Detre JA, Alsop DC. Experimental design and the relative sensitivity of BOLD and perfusion fMRI. NeuroImage 2002; in press.

Aguirre GK, Zarahn E, D'Esposito M. The variability of human BOLD hemodynamic responses. NeuroImage 1998; 8: 360-9.

Binder JR, Swanson SJ, Hammeke TA. Determination of language dominance using functional MRI: a comparison with the WADA test. Neurology 1996; 46: 978-84.

Boynton G, Engel S, Glover G, Heeger D. Linear systems analysis of functional magnetic resonance imaging in human V1. The Journal of Neuroscience 1996; 16: 4207-21.

Detre JA, Alsop DC. Perfusion fMRI with arterial spin labeling. In: Bandettini PA and Moonen C, editors. Functional MRI. Berlin: Springer Verlag, 1999: 47-62.

Engel S, Zhang X, Wandell B. Colour tuning in human visual cortex measured with functional magnetic resonance imaging. Nature 1997; 388: 68-71.

Feeney DM, Baron J-C. Diaschisis. Stroke 1986; 17: 817-30.

Friston KJ, Price CJ, Fletcher P, Moore C, Frackowiak RSJ, Dolan RJ. The trouble with cognitive subtraction. NeuroImage 1996; 4: 97-104.

Friston KJ, Zarahn E, Josephs O, Henson RNA, Dale AM. Stochastic designs in event-related fMRI. Neuroimage 1999; 10: 607-19.

Jacobson M. Developmental Neurobiology. Vol 256-258. New York: Plenum Press, 1978.

Jarrard LE. On the role of the hippocampus in learning and memory in the rat. Behavioral and Neural Biology 1993; 60: 9-26.

Jayakar P, Bernal B, Medina LS, Altman N. False lateralization of language cortex on functional MRI after a cluster of focal seizures. Neurology 2002; 58: 490-2.

Jezzard P, Song A. Technical foundations and pitfalls of clinical fMRI. NeuroImage 1996; 4: S63-S75.

Leniger-Follert E, Hossmann K-A. Simultaneous measurements of microflow and evoked potentials in the somatomotor cortex of the cat brain during specific sensory activation. Pflugers Archive 1979; 380: 85-9.

Logothetis NK, Pauls J, Augath M, Trinath T, Oeltermann A. Neurophysiological investigation of the basis of the fMRI signal. Nature 2001; 412: 150-7.

Malonek D, Grinvald A. Interactions between electrical activity and cortical microcirculation revealed by imaging spectroscopy: implications for functional brain mapping. Science 1996; 272: 551-4.

Menon RS, Luknowsky DC, Gati JC. Mental chronometry using latency-resolved functional MRI. Proceedings national academy of science, USA 1998; 95: 10902-7.

Moonen CTW, Bandettini PA, editors. Functional MRI. Berlin: Springer Verlag, 1999.

Ogawa S, Menon RS, Tank DW, Kim SG, Merkle H, Ellermann JM, et al. Functional brain mapping by blood oxygenation level-dependent contrast magnetic resonance imaging. A comparison of signal characteristics with a biophysical model. Biophysical Journal 1993; 64: 803-12.

Posner MI, Petersen SE, Fox PT, Raichle ME. Localization of cognitive operations in the human brain. Science 1988; 24: 1627-31.

Price CJ, Friston KJ. Cognitive conjunctions: a new experimental design for fMRI. NeuroImage 1997; 5: 261-70.

Price CJ, Friston KJ. Scanning patients with tasks they can perform. Human Brain Mapping 1999; 8: 102-8.

Rees G, Frith CD, Lavie N. Modulating irrelevant motion perception by varying attentional load in an unrelated task. Science 1997; 278: 1616-9.

Sternberg S. The discovery of processing stages: extensions of Donder's method. Acta Psychologica 1969; 30: 276-315.

Worsley KJ, Friston KJ. The analysis of fMRI time-series revisted-again. NeuroImage 1995; 2: 173-82.

Zarahn E, Aguirre GK, D'Esposito M. A trial-based experimental design for fMRI. NeuroImage 1997; 6: 122-38.

Tables

Task	Cognitive Process
determine if a stimulus is the same as one seen several seconds ago	working memory
match rotated figures	mental rotation
generate a verb for a supplied noun	semantic recall

Table 1: examples of cognitive processes and tasks purported to evoke them

Figures

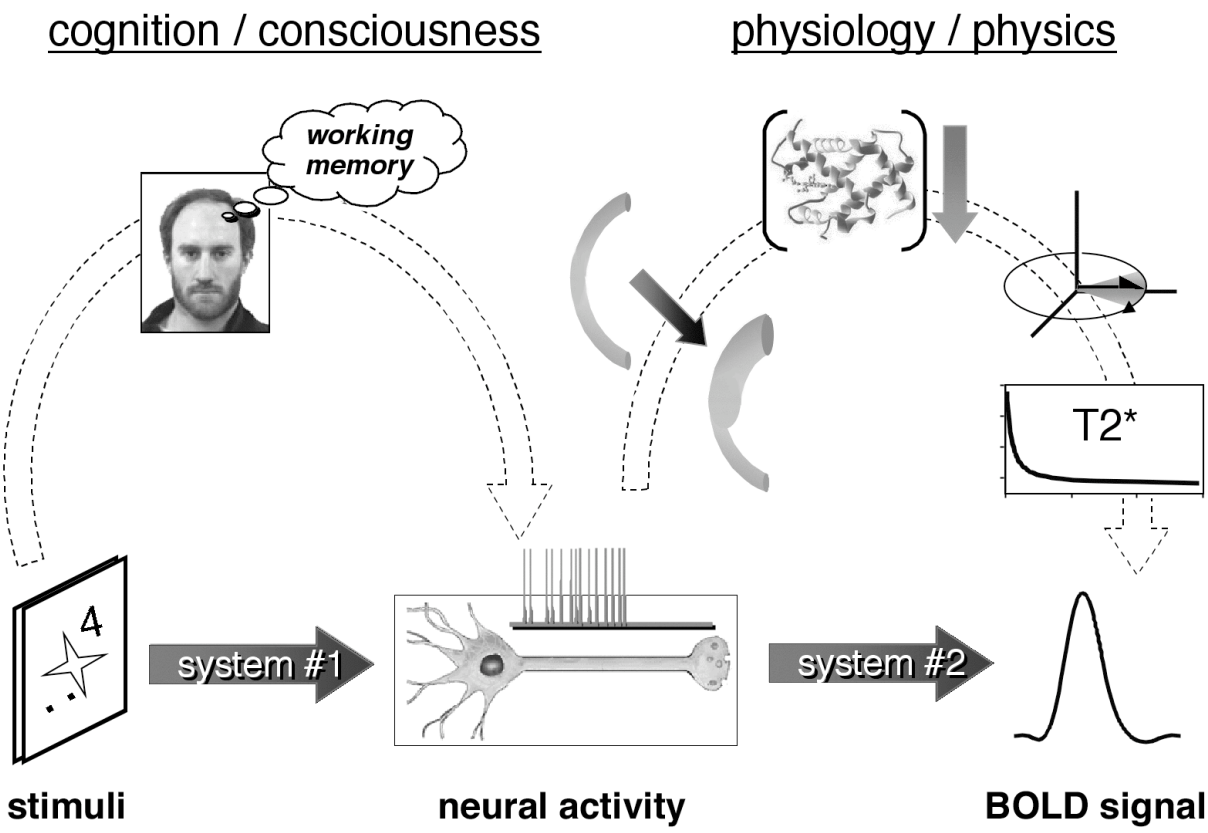

Figure 1: Depiction of the "two systems" model of BOLD fMRI experiments. The left side of the figure represents the system that mediates the transformation of stimuli and instructions in to neural activity. The cognitive processes that mediate this transformation are typically the subject of study of neuropsychology experiments. The right side of the figure depicts the system that transforms neural activity into BOLD fMRI signal. Following along the curved arrow on the right, the chain of events that follows an increase in neural activity includes dilatation of local vasculature, a decrease in the concentration of deoxyhemoglobin, an alteration in local magnetic field inhomogeneity, impacting T2* signal and ultimately BOLD fMRI signal. As is described in the text, the complicated sequence of events that take place under the "system #2" label can be efficiently modeled by a linear system.

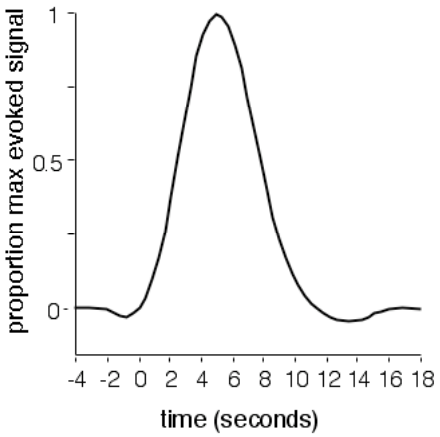

Figure 2: An average, across subject hemodynamic response function. The brief period of neural activity that produced this signal change took place at time zero.

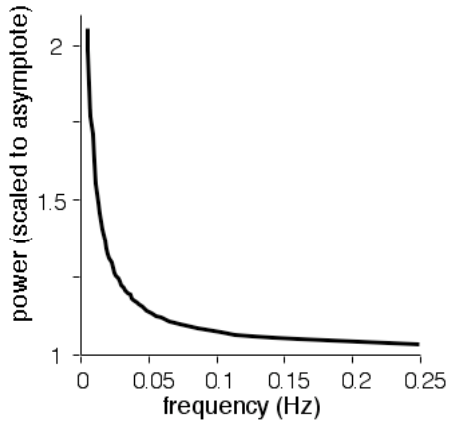

Figure 3: The power spectrum representation of the "pink," or "1/frequency" noise present in BOLD fMRI data under the null-hypothesis. Data that is composed of independent observations over time is termed "white" noise, and would have a flat line in this representation.