Brain Imaging and Cognitive Neuroscience

Toward Strong Inference in Attributing Function to Structure

Martin Sarter, Gary G. Berntson, and John T. Cacioppo
The Ohio State University

Cognitive neuroscience has emerged from the neurosciences and cognitive psychology as a scientific discipline that aims at the determination of "how brain function gives rise to mental activity" (S. M. Kosslyn & L. M. Shin, 1992, p. 146). While research in cognitive neuroscience combines many levels of neuroscientific and psychological analyses, modern imaging techniques that monitor brain activity during behavioral or cognitive operations have significantly contributed to the emergence of this discipline. The conclusions deduced from these studies are inherently localizationistic in nature; in other words, they describe cognitive functions as being localized in focal brain regions (brain activity in a defined brain region, Φ, is involved in specific cognitive function, Ψ). A broad discussion about the virtues and limitations of such conclusions may help avoid the emergence of a mentalistic localizationism (i.e., the attribution of mentalistic concepts such as happiness, morality, or consciousness to brain structure) and illustrates the importance of a convergence with information generated by different research strategies (such as, for example, evidence generated by studies in which the effects of experimental manipulations of local neuronal processes on cognitive functions are assessed). Progress in capitalizing on brain-imaging studies to investigate questions of the form "brain structure or event Φ is associated with cognitive function Ψ" may be impeded because of the way in which inferences are typically formulated in the brain imaging literature. A conceptual framework to advance the interpretation of data describing the relationships between cognitive phenomena and brain structure activity is provided.

We praise the "lifetime of study," but in dozens of cases, in every field, what was needed was not a lifetime but rather a few short months or weeks of analytic inductive inference. . . We speak piously of taking measurements and making small studies that will "add another brick to the temple of science." Most such bricks just lie around the brickyard. (Platt, 1964, p. 351)

Methods for stereogeometric functional brain imaging have become a reality over the past two decades, and their recent application to studies of cognitive function offers considerable promise for the interdisciplinary study of brain–behavior relationships. There is an intuitive, almost seductive appeal to the view that a proper understanding of the brain may be couched in terms of the selective activation of its parts during particular cognitive activities. Although lesion studies also generate evidence relevant to brain–behavior relationships (e.g., the absence of area Φ results in impairments in cognitive function Ψ; therefore, area Φ is involved in cognitive function Ψ), monitoring changes in activity in intact brain regions is generally thought to provide a more valid basis for conclusions about the normal cognitive functions of an area. It comes as no surprise, therefore, that the development of modern brain-imaging techniques has been highly instrumental in the rapid development and maturation of cognitive neuroscience, in which one stated objective is to "correlate specific aspects of information processing with specific brain regions" (Petersen, Fietz, & Corbetta, 1992, p. 217; see also Kandel & Squire, 1992; Posner & Raichle, 1994).

Localizationist conclusions ("cognitive function Ψ is localized in brain area Φ") derived from imaging studies have had significant impact on public views about the functioning of the brain, as revealed in popular accounts of "the brain's moral center," "pinpointing chess moves in the brain" (Blakeslee, 1994, p. B1), or "how the brain computes tears and laughter" (Goleman, 1995, p. B9). More important, the localizationistic perspective as derived from imaging studies is impacting brain–behavior concepts in diverse areas related to cognitive neuroscience, including psychology (e.g., Chabris & Hamilton, 1992) and psychiatry (e.g., George et al., 1995; Liddle, 1992). This perspective is also illustrated by the questions increasingly posited in neuropsychology: "How does regional brain activity of a mathematician differ from that of an architect? . . . Which brain regions does a lawyer activate when preparing or presenting a case?" (Gur & Gur, 1991, p. 432). This line of thinking is not merely heuristic but may shake the very conception of neurobehavioral organization, as reflected, for example, in the
query "Could we train individuals to activate appropriate brain regions and thereby enhance their abilities?" (Gur & Gur, 1991, p. 432). Imaging studies have become a major force in national neuroscience policy and have emerged as a basis for the definition of programmatic research goals. An example can be found in one of the chief goals of the Human Brain Project, a substantive cooperative effort of multiple federal agencies to "provide tools for the acquisition, storage and transmission of 3-dimensional images of the human brain's structure and function" (Healy, 1992, p. 1).

Because of the enormous significance of imaging techniques for research in cognitive neuroscience, a broad discussion about the potential and limitations of conclusions derived from imaging studies may be worthwhile. Although cautionary voices have been heard (Gross, 1985; Sergent, 1994), such a debate has not emerged as a salient feature of this literature (Druckman & Lacey, 1989; Kandel & Squire, 1992; Kosslyn & Shin, 1992; Petersen et al., 1992; Posner, Petersen, Fox, & Raichle, 1988), nor has prior discussion thoroughly articulated the formal structure of scientific inference that underlies much of the brain imaging research on cognitive functions. A broader conceptual framework may be helpful to appropriately view and evaluate conclusions such as "specialized neuronal systems in the right superior temporal cortex participate in perceptual analysis of melodies" (Zatorre, Evans, & Meyer, 1994, p. 1908), or "right prefrontal cortex is an important structure for the emotional responses concerning human male sexuality" (Thihonen et al., 1994, p. 243). Such a framework may also reveal or highlight the conditions required to achieve a convergence of the "top-down" approaches characteristic of cognitive neuroscience with the more "bottom-up" strategies of behavioral neuroscience (see also Churchland, 1986) and yield a more optimal multilevel integrative analysis of neurobehavioral functions (Cacioppo & Berntson, 1992a). We begin by reviewing briefly the history of localization and its current status in imaging studies of brain–behavior relationships.

### Brain Imaging and Cerebral Localizationism: Historical Perspective

The potential use of brain imaging for the study of cognitive functions derives from the explicit or implicit assumption that cognitive operations are localizable to focal brain regions or systems. It could be argued that functions can never be localized, as functions do not represent properties of neuronal circuits but rather manifestations of the operations or outputs of these circuits. Consequently, when one speaks of localization of function it is important to understand that this term represents a shorthand notation of the fact that functions may be causally related to operations of specific neuronal circuits. Functions, in fact, one step removed from outputs, reflect the consequences or significance of outputs. It is in this sense that we use the phrase *functional localization*.

The extent to which particular functions can be localized to specific brain areas has been a subject of considerable historical debate. The first example of localization of function in a brain structure has been traced to Julien Jean César Legallois in 1812, who identified a focal region of the medulla essential for respiration (Finger, 1994). Shortly thereafter, Sir Charles Bell (1811/1936) and François Magendie (1822) demonstrated in lesion experiments on dogs that the spinal roots were functionally and anatomically distinct, with sensory and motor functions ascribed to dorsal and ventral roots, respectively. These observations led Bell to suggest that sensory and motor functions also may be localized in the brain—a hypothesis that has been clearly confirmed. A more precise localization of cortical sensory and motor areas was provided by the early stimulation and lesion studies of Fritsch and Hitzig (1870) and Ferrier (1876), and subsequent studies further documented a precise topographic representation in cortical sensory and motor areas. Despite the emerging evidence on localization of sensory and motor areas of the cortex, questions as to the local-
Gary G. Berntson

izability of more complex cognitive functions continue to the present day.

Early in the 19th century, Franz Joseph Gall, who was highly regarded for his careful anatomical studies (Zola-Morgan, 1995), proposed the theory that higher mental functions were subserved by specific and distinct areas of the cerebrum. The most controversial aspect of Gall's concepts was that the development of these brain regions (and mental functions) was marked by features of the overlying cranium. The evidence marshaled by Gall in support of his theory came principally from individuals who were characterized either by unusual crania or unusual behaviors (Finger, 1994). This led to a spirited controversy between the cerebral holism of Pierre Flourens and the ultralocalizationism (phrenology) of Francis Gall (Flechsig, 1896; Phillips, Zeki, & Barlow, 1984; Plum & Volpe, 1987; Zülch, 1976). This debate was featured at the 1881 meeting of the International Medical Congress in London, with Ferrier advancing the localizationist view, and Goltz challenging at least the strongest form of that position (see Phillips et al., 1984).

The localizationist controversy has continued throughout this century. Despite the proliferating demonstrations of apparent localization of cognitive functions and language processes in the cortex, many researchers such as Karl Lashley (1929) argued stridently for a functional "equipotentiality" of the neocortex in learning and memory processes. The debate over brain localization was at the core of the heated controversies of the 1960s and 1970s over the use, feasibility, and ethics of psychosurgery (e.g., Valenstein, 1973, 1980). This debate clearly continues (e.g., Farah, 1994), as reflected in Mesulam's (1994) observation, "The field of cognitive neuroscience is supposed to be engaged in an epic struggle between the forces of localization (or centrism) and those of equipotentiality (or holism)" (p. 74).

Squire (1987) suggested that the relevant issue is not localizationism versus antilocalizationism per se, but rather "concerns the size of the ensembles, i.e., the size of the functional unit within which information is equivalently and statistically distributed" (p. 319). According to this view, the neuronal substrates of cognitive operations are ultimately localizable, on the level of individual neurons, groups of neurons, or neuronal networks. Hence, the apparent limitations of current imaging studies to yield specific and definitive localizationistic conclusions may be attributable to the relatively low resolution of current technology. The hope is that technological progress will allow monitoring of the activity of appropriately small functional units over relatively short periods of time (M. S. Cohen & Bookheimer, 1994; see also Sereno et al., 1995) and thus mollify residual concerns over the limitations of localizationism.

Cerebral Localization of Function: Methodological and Logical Issues

Although a refinement in the spatial and temporal resolution of imaging methods would be of unquestioned benefit, more significant issues relate to the fundamental nature of structure–function relationships and the requisite logic for deriving inferences concerning these relationships from imaging data. These issues are largely separate from the concern over resolution, and the mere decrement in the size of the functional unit does not offer immunity against inappropriate conclusions. An important question that arises in any attempt at brain localization is what, precisely, is localized? This may not be particularly problematic for simple spinal reflexes, where a reasonable isomorphism can be shown between concepts of the operations of neural circuits (e.g., patterns of synaptic contact, synaptic potentials, neuromuscular activation, etc.) and conceptions of functional organizations (e.g., pain-withdrawal reflex, stepping reflexes, etc.). Adding complexity to the issue of localization, however, is the fact that cognitive functions are often based on multiple processes (sensory–perceptual analysis, learning, memory processing). In fact, each of these multiple processes in turn may be composed of many distinct neuronal operations that may show varying degrees of localizability. In attempting to map complex functions onto complex structures, there is a considerable likelihood that concepts and models at the functional (cognitive) level, although they may overlap, may not be isomorphic with concepts and models of neural systems or processes. That is, the formal structure of the cognitive models may not correspond to the multiple brain systems, mechanisms, or neuronal codes that underlie the cognitive functions. A lack of function–structure isomorphism can hamper attempts to understand brain–behavioral relationships. Any meaningful neurobiological model of visual perception, for example, must recognize the differential encoding of orientation, movement, and color by subcomponents of the geniculostriate path-
that are unique to the stimulated task state (Fox, Mintun, Reiman, & Raichle, 1988). The general limitations of, caveats to, and benefits of the subtractive method in the analysis of physiological data have been discussed elsewhere (e.g., Cacioppo & Petty, 1986; Mintun, Fox, & Raichle, 1989; Posner et al., 1988). Prior to paired image subtraction, however, the measures in each image must be standardized in some fashion (e.g., unitized) to equate for overall differences in the measured activation across images. Although this standardization can dramatically enhance the signal–noise properties of these data, it can also introduce apparent differences between images that are artifactual.

In a simplified example, consider a case in which a control task (e.g., gazing at a fixation point on a monitor) produced no activation in the cerebellum and a graded distribution of activation from frontal to occipital regions, and a stimulated task (e.g., passively viewing lines and contours) again produced no activation in the cerebellum, the same gradient of activation from frontal to occipital regions, but higher absolute measures from these activated regions. If each image is unitized by dividing each pixel intensity by the total intensity of the image, the cerebellum can appear in the subtracted image as a structure in which activity was diminished. This finding would represent a spurious consequence of data transformations (see Levey, 1980).

Beyond these technical and methodological issues lie more fundamental limitations in the ability of brain imaging approaches to accurately reveal existing functional localization. Even for functions that are in fact localized to specific neural circuits, these circuits may (a) be diffusely organized or widely distributed; (b) anatomically overlap, or even share common neuronal elements with circuits mediating different functions; or (c) perform different functions depending on the patterns of input–activation associated with different cognitive states or contexts. These possibilities would clearly complicate efforts to elucidate the cerebral localization of functions.

Methodological and Interpretative Issues

One obstacle to inferences about the cognitive significance of brain imaging data stems from the fact that advances in the comprehensive representation and analysis of complex physiological signals such as brain images have lagged behind advances in signal acquisition. Brain images that are based on positron emission tomography (PET), for example, may provide an accurate representation of the spatial distribution of the radionuclide at selected planes. Inferences concerning neurocognitive functions, however, require assumptions about the relationships between radio-label density, local metabolism, regional blood flow, and the pattern of neuronal activity. Although many of these assumptions may be reasonable, others may be more problematic. Increasing the local activity of excitatory or inhibitory interneurons in a given area, for example, may yield a similar radio-emission signal while having divergent functional implications (Posner & Raichle, 1994).

An additional issue derives from the fact that PET images of functional activation are generally based on paired image subtraction (Raichle, 1994). Two tasks (a control and stimulated task) are developed that, theoretically, differ only in terms of the cognitive operations of interest. The resulting images derived from the control task are subtracted from those obtained during the stimulated task to obtain an image that presumably reflects differences in activation between the two tasks. The resulting difference image is interpreted as revealing those areas of the brain concerned with the cognitive operations that are unique to the stimulated task state (Fox, Mintun, Reiman, & Raichle, 1988).
Activation of diffusely organized circuits (or components of circuits), for example, may not yield a sufficiently intense signal to be differentiated from noise or from activity patterns of control tasks. This could lead to a failure to detect a localized function. Alternatively, a local, detected region within a larger undetected area of activation could lead to an overestimate of the degree of functional localization. Similarly, the existence of functionally distinct but overlapping circuits could lead to an underestimate of the degree of functional localization, as the overlapping area would be activated during multiple cognitive or behavioral contexts. The latter appears to account at least in part for the historical controversy, discussed above, over the motivational specificity of hypothalamic and limbic circuits (Berntson & Beattie, 1975; Roberts, 1969).

Perhaps the more problematic possibility, for which there is ample precedent, is that some central circuits may have differential and overlapping functions depending on the pattern of activation. A helpful example is provided by the research on the functions of "state-setting" systems (Mesulam, 1990), particularly the attentional functions mediated by cortical cholinergic and noradrenergic afferents. Research in this area has resulted in fairly specific hypotheses about the attentional processes that depend on the integrity of cortical cholinergic and noradrenergic afferents and about the key role of attentional dysfunctions on the basis of aberrations in activity of these systems in major neuropsychiatric disorders (e.g., Robbins & Everitt, 1987; Sarter, 1994; Sarter & Bruno, 1994). Whereas imaging studies have pointed to cortical areas in frontal, temporal, and parietal lobes that may be involved in various types of attention (R. M. Cohen, Semple, Gross, King, & Nordahl, 1992; Corbetta, Miezin, Dobmeyer, Shulman, & Petersen, 1991; Grossman, Crino, Reivich, Stern, & Hurtig, 1992; Pardo, Fox, & Raichle, 1991; Posner & Petersen, 1990; Posner et al., 1988), more "bottom-up" approaches aim at the determination of the specific role of acetylcholine in different cortical areas in attentional functions (e.g., Metherate & Weinberger, 1990). An understanding of the role of the major afferent projections of cortical areas in attention offers considerable promise for understanding the functional deficits in schizophrenia and dementia and for the development of intervention strategies. This effort will benefit from a true convergence of evidence from imaging studies in humans and from animal studies on the cognitive effects of manipulations of the activity of cortical inputs (e.g., Sarter & Bruno, 1994). Imaging studies may continue to relate sustained attention to prefrontal and superior parietal areas (Pardo et al., 1991), for example, and decreases in the activity in these areas may indeed be associated with the cognitive decline in dementia (e.g., Mielke, Herholz, Grond, Kessler, & Heiss, 1994). In isolation, however, the imaging approach is unlikely to reveal the precise functional contribution of individual afferent systems of these areas and, more generally, cannot discriminate different neuronal activity patterns leading to identical signal levels. Note that this example is not meant to contrast or even prioritize the heuristic power of the two levels of analysis; rather, it serves to illustrate the significance of an integration of relevant evidence from the basic neurosciences in the interpretation of imaging data to specify the neuronal networks and mechanisms mediating the components of attention (Mesulam, 1990). As is discussed below, the means by which top-down and bottom-up information can be integrated determines the formulation of structure-function relationships.

**Toward Strong Inference in the Interpretation of Brain–Behavior Relationships**

There is an intuitive appeal to the view that a proper understanding of the neural substrates of cognition may be couched in terms of the selective activation of regions of the brain during particular cognitive activities. Although progress has been made, the manner in which many inferences are drawn about the cognitive significance of localized brain activity remains problematic. In this section, we consider the formal structure of inferences to be drawn from brain-imaging data and offer an alternative framework for deriving meaningful interpretations of these data.

A major goal of cognitive neuroscience can be conceptualized in the following relationship:

$$\Psi = f(\Phi),$$

where $\Psi$ represents specific cognitive operations or functions and $\Phi$ represents the underlying neural substrates and processes (Cacioppo & Tassinary, 1990).

Given the relative youth of the field of brain imaging, it should not be a surprise that studies can be found in which an experiment is performed, observations are made, and these observations act as a starting point for a series of internally consistent propositions, culminating in a general conclusion (model) that is consistent with the empirical observations in the experiment. For instance, investigators might vary cognitive operations by contrasting tasks that are thought to differ only in one or more cognitive functions ($\Psi$) and then isolate the differences between the associated brain images ($\Phi$). These data are then interpreted as showing that brain structure (or event) $\Phi$ is associated with cognitive function $\Psi$; these data are also treated as revealing much the same information that would have been obtained had brain structure (or event) $\Phi$ been stimulated or ablated and a consequent change in cognitive function $\Psi$ been observed. This form of interpretation reflects the explicit assumption that there is a fundamental localizability of specific cognitive operations and the implicit assumption that there is an isomorphism between $\Phi$ and $\Psi$.

It is of interest to note that prior to the development of hypothetico-deductive logic by Sir Francis Bacon in 1620, scientists were content with performing an experiment and recording their observations. As in some contemporary brain-imaging research, these observations would represent a starting point for a series of internally consistent propositions, culminating in a general conclusion. Bacon's important contribution was the rigorous application of a special kind of inductive reasoning pro-
ceeding from the accumulation of empirical observations to the general conclusion (see Brazier, 1959). The components of Bacon's inductive reasoning are now common knowledge and a cornerstone to many, though not all, scientific disciplines (Platt, 1964): (a) Identify a phenomenon of interest, (b) distinguish two or more hypotheses about the phenomenon, (c) devise a set of conditions with alternative possible outcomes excluding one or more of these hypotheses, (d) establish the conditions and collect observations while minimizing measurement error and artifacts, (e) replicate conceptually, and (f) eliminate hypotheses on the basis of discrepancies between the predicted and the observed outcomes. This hypothesis-driven research approach is of the form not-$\Psi = f(\text{not-}\Phi)$; that is, it seeks to reject the alternative hypothesis.

Causal hypotheses regarding a specific brain structure or process ($\Phi$) underlying a cognitive operation ($\Psi$) are of the form $\Psi = f(\Phi)$ and necessarily imply that $\Phi$ is always followed by $\Psi$ but do not necessarily imply that $\Psi$ is always preceded by $\Phi$. Furthermore, brain events ($\Phi$) may be of interest to the extent that they index a cognitive operation or state so that the inferences are of the form $\Psi = f(\Phi)$ rather than not-$\Psi = f(\text{not-}\Phi)$. Thus, an ultimate aim of brain imaging in cognitive neuroscience can generally be specified by the conditional probability of $\Psi$, given $\Phi$:

$$P(\Psi/\Phi) = 1. \quad (2)$$

In contrast, however, the typical structure of the imaging study can be represented by the conditional probability of $\Phi$, given $\Psi$:

$$P(\Phi/\Psi) = x. \quad (3)$$

That is, brain imaging techniques (e.g., magnetoencephalography [MEG], positron emission tomography [PET], functional magnetic resonance imaging [fMRI], electroencephalography [EEG]) provide information about $\Phi$ as a function of $\Psi$, but the conditional probabilities of Equations 2 and 3 are equivalent: $P(\Psi/\Phi) = P(\Phi/\Psi)$ if and only if there is a 1:1 relationship between brain structure (or event) $\Phi$ and cognitive functions or operations $\Psi$ (i.e., if $x = 1$ in Equation 3).

Most of the early discoveries of functional localizationism that have had lasting impact (e.g., Legallois' research on the medullary respiratory center; Fritsch and Hitzig's research on the motor cortex; Munk's research on the visual cortex) involved the direct electrical stimulation or ablation of specific areas of the brain of animals (i.e., $P(\Psi/\Phi)$). The evidence for Gall's theory, in contrast, was based primarily on relating cranial features to extreme behaviors (i.e., $P(\Phi/\Psi)$) and was interpreted to mean that individuals with these cranial features were destined toward these behaviors ($P(\Phi/\Psi)$). The latter structure of scientific inference is not necessarily misleading if $P(\Phi/\Psi) = P(\Psi/\Phi)$, which, of course, was not the case with respect to Gall's conclusions.

When general conclusions in brain-imaging studies are reached about the involvement (i.e., causal role) of a brain structure or event in a cognitive function, it is assumed, at least implicitly, that there is a 1:1 relationship. That is, the empirical observation that $P(\Psi/\Phi)$ is treated as equivalent to $P(\Phi/\Psi)$. It is important that these interpretations be based on more than mere consistency with the imaging data, however. Ideally, evidence would also be presented demonstrating that the $P(\text{not-}\Psi/\Phi) = 0$ in the assessment context (in which case $\Phi$ would be a marker of $\Psi$) or across contexts (in which case the relationship between $\Phi$ and $\Psi$ would be invariant). Strong inferences would then be possible (Cacioppo & Tassinary, 1990). In the absence of such evidence, progress in cognitive neuroscience may be fostered if the interpretations of brain-imaging data were tempered to reflect the lack of information about the specificity of the neurocognitive relationship, in other words, about the validity of the assumption that $P(\Psi/\Phi) = P(\Phi/\Psi)$.

Approaches such as stimulation and ablation studies and brain-imaging research provide complementary rather than redundant information about the relationship between brain structures (or events) and cognitive functions. This is because stimulation and ablation studies bear on the relationship $P(\Psi/\Phi)$, whereas brain-imaging studies provide information about $P(\Phi/\Psi)$. Despite the formal parallelism between the expression $P(\Psi/\Phi)$ and $P(\Phi/\Psi)$, there is a fundamental asymmetry in the heuristic power of studies aimed at the demonstration of $P(\Psi/\Phi)$ versus $P(\Phi/\Psi)$. The causal role of $\Phi$ in process $\Psi$ can be examined in a straightforward fashion by direct experimental manipulation of the brain. The loss of cognitive function by inactivation of neuronal processes (by mechanical, thermal, or neurochemical means) can serve to establish $\Phi$ as necessary for function $\Psi$. Moreover, addressing a more complex avenue of research, facilitation of cognitive processes (e.g., attention or memory) by electrical or neurochemical brain activation (e.g., Sarter et al., in press) can further establish that $\Phi$ is a sufficient condition for facilitation of process $\Psi$.

Although both relationships $P(\Psi/\Phi)$ and $P(\Phi/\Psi)$ can be studied experimentally, the complexity implied in studies aimed at the demonstration of $P(\Psi/\Phi)$ is based on the fact that experimental alteration of cognitive function $\Psi$ necessarily alters brain activities $\Phi_1, \Phi_2, \ldots$ (including those that underlie $\Psi$ [say $\Phi_1$]). Thus, although a psychological context may yield concurrent manifestation in $\Psi$ and $\Phi_1$, the causal linkage remains in question, as an alternate (even undetected) brain event $\Phi_2$ could be causally mediating both $\Psi$ and $\Phi_1$. In this case, excluding $\Phi_1$ as a necessary or sufficient condition for the observed $\Psi$ may require a tedious and exhaustive evaluation of the alternative hypothesis not-$\Psi = f(\text{not-}\Phi_1)$. Note that this conclusion does not challenge the importance and use of evidence generated by imaging methods but points to fundamental limitations in the strength of the inference deduced from experimental approaches aimed solely at

\[ The \text{interested reader may wish to see Cacioppo and Berntson (1992b) for a discussion of the distinction between parallel and convergent multiple determinism.} \]
the demonstration of $P(\Psi/\Phi)$. The integration of methods and data from bottom-up and top-down approaches provides a means of circumventing some of the thornier interpretive problems of either approach alone and thereby permits strong inferences in cognitive neuroscience.\footnote{Research on the role of the nucleus pulvinaris of the thalamus in attention exemplifies the importance of this convergence of approaches. Studies using PET measurements of glucose incorporation revealed that object identification processes that demand selective attention increase pulvinar activity (i.e., these studies provided information in the form $\Phi = \{\Psi\}$; LaBerge & Buchsbaum, 1990). Evidence from studies in various species, including humans, shows that lesions of this area affect visual spatial attention (i.e., information in the form $\Psi = \{\Phi\}$; e.g., Rafal & Posner, 1987). Furthermore, the specific neuronal coding is beginning to be studied. Single-unit recordings in behaving subhuman primates reveal that increases in pulvinar activity indicate accurate eye movements to the intended target and suppression of inaccurate eye movements (i.e., $\Phi = \{\Psi\}$; Robinson, McCurkin, & Kertzman, 1990). The converging evidence from these studies suggests that, in this case, the assumption that $P(\Phi/\Psi) = P(\Psi/\Phi)$ is valid. Thus, strong inferences from imaging data may be possible, provided that additional evidence in the form of $P(\Psi/\Phi) = P(\Phi/\Psi)$ will be available. This information, together with further studies of the specific neuronal processes, promises to offer fundamental insights into the neuronal and cognitive mechanisms of visual attention and the role of the pulvinar nucleus in these processes.}

\textbf{An Illustration}

It might be useful to illustrate some of these points using a simple physical metaphor in which—unlike in cognitive neuroscience at present—the bases of a multiply determined outcome are known. Briefly, let $\Phi$ represent the heater and $\Psi$ the temperature in a house. In the context of cognitive neuroscience, the heater parallels a neuronal mechanism and the temperature represents the cognitive manifestation of the operation of this mechanism. Although the heater and the temperature are conceptually distinct, the operation of the heater represents a physical basis for the temperature in the house. Thus, $\Psi = f(\Phi)$. A bottom-up approach (i.e., $P(\Psi/\Phi)$) makes clear certain details about the relationship between $\Psi$ and $\Phi$, whereas a top-down approach (i.e., $P(\Phi/\Psi)$) clarifies others. For instance, when the activity of the heater is manipulated (i.e., $\Phi$ is stimulated or lesioned), a change in the temperature in the house ($\Psi$) results. This represents a bottom-up approach to investigating the neuronal substrates of cognitive phenomena. The fact that manipulating the activity of the heater produces a change in the temperature in the house can be expressed as $P(\Psi/\Phi) > 0$. Note that the $P(\Psi/\Phi)$ need not equal 1 for $\Phi$ to be a physical substrate of $\Psi$. This is because, in our illustration, there are other physical mechanisms that can affect the temperature in the house ($\Psi$), such as the outside temperature ($\Phi_1$) and the amount of direct sunlight in the house ($\Phi_2$). That is, there is a lack of complete isomorphism specifiable, at least initially, between the functional dimension ($\Psi$) and a physical basis ($\Phi$).

In any given context, the temperature in the house may be influenced by any or all of these physical mechanisms. If the outside temperature or the amount of direct sunlight happens to vary when the heater is activated, then the temperature may not covary perfectly with the activation of the heater (i.e., $P(\Psi/\Phi) < 1$) even though the temperature is, at least in part, a function of the operation of the heater (i.e., $P(\Psi/\Phi) > 0$). If the outside temperature and amount of direct sunlight are constant or are perfectly correlated with the activation of the heater, then the temperature in the house and the activity of the heater may covary perfectly (i.e., $P(\Psi/\Phi) = 1$). In the context of cognitive neuroscience, this is analogous to a brain lesion study accounting for some of the variance ($P(\Psi/\Phi) > 0$) or all of the variance ($P(\Psi/\Phi) = 1$) in cognitive measure in the study. The latter result does not imply the lesioned brain region is a necessary component just as the temperature in the house covarying perfectly with the temperature in the house does not mean necessarily that there are not other physical mechanisms that may also influence the temperature. Thus, as long as $P(\Psi/\Phi) > 0$, $\Phi$ could be considered a predictor (or component) of $\Psi$; that $P(\Psi/\Phi) = 1$ does not imply that $\Phi$ is the only or a necessary cause of $\Psi$.

The asymmetry between $P(\Psi/\Phi)$ and $P(\Phi/\Psi)$ and the interpretive problems that may result when simply assuming $P(\Psi/\Phi) = P(\Phi/\Psi)$ are also evident in this metaphor. As outlined above, the former term represents variations in temperature in the house given variations in the activity of the heater, whereas $P(\Phi/\Psi)$ represents the activity of the heater given variations in the temperature in the house. Although one would expect to find $P(\Psi/\Phi) > 0$ in some contexts, the fact that the temperature in the house increases reliably when the heater is activated does not necessarily imply that changes in the temperature in the house are associated with variations in the activity of the heater. In the winter months, changes in the temperature in the house may be associated with corresponding changes in the activity of the heater. In another context (e.g., the summer months), however, the heater may be uniformly inactive but housing temperature continues to vary because of the operation of other physical factors (e.g., outside temperature, $\Phi_1$; exposure to direct sunlight, $\Phi_2$). Thus, the finding that $P(\Psi/\Phi) = 0$ does not mean $\Phi$ has no role in $\Psi$, only that $\Phi$ has no role in $\Psi$ in that context. In the context of brain-imaging studies, areas that are not found to become active as a function of a cognitive operation may nevertheless be part of a physical substrate for that cognitive operation (just as a heater may remain a part of the physical mechanism for the temperature in a house).

The preceding example illustrates why one would not want to exclude a brain area as potentially relevant to a cognitive operation on the basis of the area not being activated in a brain image as a function of the cognitive operation. The converse also holds—that is, a brain area that is activated as a function of a cognitive operation may or may not contribute meaningfully to the production of the cognitive operation. Consider an LED on-off indicator on the heater (which we will call $\Phi_3$) that illuminates when the heater ($\Psi$) is operating. In this case, the $P(\Phi/\Psi) = P(\Phi_3/\Psi) > 0$. That is, the LED represents a physical element that would show the same covariation in any given context, the temperature in the house may be influenced by any or all of these physical mechanisms. If the outside temperature or the amount of direct sunlight happens to vary when the heater is activated, then the temperature may not covary perfectly with the activation of the heater (i.e., $P(\Psi/\Phi) < 1$) even though the temperature is, at least in part, a function of the operation of the heater (i.e., $P(\Psi/\Phi) > 0$). If the outside temperature and amount of direct sunlight are constant or are perfectly correlated with the activation of the heater, then the temperature in the house and the activity of the heater may covary perfectly (i.e., $P(\Psi/\Phi) = 1$). In the context of cognitive neuroscience, this is analogous to a brain lesion study accounting for some of the variance ($P(\Psi/\Phi) > 0$) or all of the variance ($P(\Psi/\Phi) = 1$) in cognitive measure in the study. The latter result does not imply the lesioned brain region is a necessary component just as the temperature in the house covarying perfectly with the temperature in the house does not mean necessarily that there are not other physical mechanisms that may also influence the temperature. Thus, as long as $P(\Psi/\Phi) > 0$, $\Phi$ could be considered a predictor (or component) of $\Psi$; that $P(\Psi/\Phi) = 1$ does not imply that $\Phi$ is the only or a necessary cause of $\Psi$.

The asymmetry between $P(\Psi/\Phi)$ and $P(\Phi/\Psi)$ and the interpretive problems that may result when simply assuming $P(\Psi/\Phi) = P(\Phi/\Psi)$ are also evident in this metaphor. As outlined above, the former term represents variations in temperature in the house given variations in the activity of the heater, whereas $P(\Phi/\Psi)$ represents the activity of the heater given variations in the temperature in the house. Although one would expect to find $P(\Psi/\Phi) > 0$ in some contexts, the fact that the temperature in the house increases reliably when the heater is activated does not necessarily imply that changes in the temperature in the house are associated with variations in the activity of the heater. In the winter months, changes in the temperature in the house may be associated with corresponding changes in the activity of the heater. In another context (e.g., the summer months), however, the heater may be uniformly inactive but housing temperature continues to vary because of the operation of other physical factors (e.g., outside temperature, $\Phi_1$; exposure to direct sunlight, $\Phi_2$). Thus, the finding that $P(\Psi/\Phi) = 0$ does not mean $\Phi$ has no role in $\Psi$, only that $\Phi$ has no role in $\Psi$ in that context. In the context of brain-imaging studies, areas that are not found to become active as a function of a cognitive operation may nevertheless be part of a physical substrate for that cognitive operation (just as a heater may remain a part of the physical mechanism for the temperature in a house).

The preceding example illustrates why one would not want to exclude a brain area as potentially relevant to a cognitive operation on the basis of the area not being activated in a brain image as a function of the cognitive operation. The converse also holds—that is, a brain area that is activated as a function of a cognitive operation may or may not contribute meaningfully to the production of the cognitive operation. Consider an LED on-off indicator on the heater (which we will call $\Phi_3$) that illuminates when the heater ($\Psi$) is operating. In this case, the $P(\Phi/\Psi) = P(\Phi_3/\Psi) > 0$. That is, the LED represents a physical element that would show the same covariation in
with the temperature in the house as would the operation of the heater as long as a top-down approach was used. When the complementary bottom-up approach was used, it became obvious that disconnecting (lesioning) the heater has effects on the temperature in the house whereas disconnecting (or directly activating) the LED has none.

The metaphor also illustrates the categorical error: the intuitively appealing notion that the organization of cognitive phenomena maps in a 1:1 fashion into the organization of the underlying neural substrates. The temperature of the house, for instance, does not map into a single “temperature center” in the house but rather is determined by several different physical mechanisms. One might argue that the problem is in how $\psi$ is being defined, that the temperature in the house is not the most useful way of conceptualizing the phenomenon. One solution may be to specify the heater-related contribution to temperature (reconceptualize the function dimension as $\psi$). Although this represents one approach, it is not the only, or even necessarily the best, solution. For instance, what if the system is designed to maintain the temperature of the house within a narrow range? The problem in such a system may not be in the definition of $\psi$ but in the full delineation of its physical basis ($\phi_1$, $\phi_2$, and interactions) to achieve a closer approximation between functional concepts and models of mechanisms. Although we anticipate that 1:1 mappings between $\phi$ and $\psi$ will be achieved ultimately, reaching this ultimate aim requires a recognition of the preliminary state of our knowledge at present and the attendant implications for strong inference. Given the complementary nature of the data from brain imaging and from direct stimulation and lesion studies, progress in cognitive neuroscience should be fostered by an integration rather than a progressive segregation of these approaches and literatures.

**Conclusion**

Whether behaviorally or cognitively induced changes in neuronal activity are determined on the level of macroscopic areas, networks, or individual neurons, brain–behavior relationships per se represent localizationist conclusions. Thus, as discussed above, localizationism in terms of attempting to correlate functions with structure is not the crucial or even primary issue. The concern is that, although structure–function relationships are established by top-down approaches, these correlations may not be validly interpreted as pointing to the “actual fundamental faculties” processed by focal brain areas or circuits (see Gross’s, 1985, discussion of Gall’s prephrenology proposals). The possibility that “one cannot be assured of arriving at the correct macrostructural description” (Farah, 1994, p. 60) may not primarily be a consequence of the specific conclusions about structure–function relationships as they are typically formulated in studies using top-down approaches but rather may be a result of the structure of scientific inference. Despite the fact that area $\phi$ becomes active during happiness, for example, this does not imply that happiness is localized to area $\phi$. An integrating analysis across levels, in which research is guided by and concepts and theories are compatible with other levels of analyses, provides the basis for a cognitive neuroscience capable of strong inferences.

**REFERENCES**


