Localization and Intrinsic Function

Charles Rathkopf

Published in *Philosophy of Science*
January, 2013

Abstract

This paper describes one style of functional analysis commonly used in the neurosciences called task-bound functional analysis. The concept of function invoked by this style of analysis is distinctive in virtue of the dependence relations it bears to transient environmental properties. It is argued that task-bound functional analysis cannot explain the presence of structural properties in nervous systems. An alternative concept of neural function is introduced that draws on the theoretical neuroscience literature, and an argument is given to show that this alternative concept of may help to overcome the explanatory limitations of task-bound functional analysis.

1 Introduction

Everyone agrees with Petersen and Fiez's 1993 oft-cited claim that "there is no tennis forehand area in the brain." However, there is presently little agreement about the grounds for this widely held belief. What precisely makes "tennis forehand" an inappropriate or unlikely functional designation? Each of the following options provides the beginning of an answer: (i) it is too specific, (ii) it is too arbitrary, and (iii) it is too unrelated to the selection pressures that shaped the development of neural function. There is, no doubt, something right in each of these suggestions. But it is difficult to turn such truisms into theories that are substantial enough to impact the ongoing scientific debates about localization. In this paper, an alternative answer is proposed which, if successful, has direct consequences not only for the philosophical analysis of function, but also for neuroscientific practice. The answer proposed is roughly this: there is no tennis forehand area in the brain because the execution of a forehand is a behavioral task, and no brain area has a behavioral task as its principal function.

This paper is comprised of three main sections. In Section 2, one style of functional analysis—which I call task-bound functional analysis—is described. Task-bound functional analysis is routinely used in the neurosciences, and some examples are provided to illustrate its use. In Section 3, an argument is presented that shows why task-bound functions are incapable of explaining the presence of structural properties in nervous systems. In Section 4, an alternative concept of neural function is introduced. It is argued that the explanatory strategy associated with this alternative concept is, by virtue of satisfying two explicit desiderata, not subject to the limitations of task-bound functional analysis. By contrasting the concept of task-bound function with the newer, alternative concept, some light is shed on the explanatory depth of certain contemporary neuro-functional hypotheses.

2 Task-Bound Functional Analysis

2.1 Context Sensitivity

Even the most celebrated cases of localized neural function remain controversial. Consider the paradigmatic localization of speech production in humans to Broca's area (Broadmann areas 44 and 45). Although speech production

certainly depends on characteristic activation patterns in Broca's area, recent research suggests at least four distinct classes of proposed functions that it also, ostensibly, realizes. These proposed functions vary along multiple theoretical dimensions, and there is substantial debate regarding both the extent to which they are compatible with one another, and whether or not any of them are correct (Grodzinsky and Santi, 2008). Why should there be such controversy about the legitimacy of even the best cases of localized neural function? The answer, I submit, lies in the observation that many neuro-functional hypotheses make essential reference to transient environmental contexts.[1]

One might think that making reference to transient environmental contexts is precisely what neuro-functional hypotheses ought to do. Indeed, in Section 2.3, I argue that such neuro-functional hypotheses play a crucial role in neuroscientific investigation. One scientific motivation for appealing to functions that essentially depend on reference to transient environmental properties is that such functions are relatively easy to relate to a lower level of description. Often, researchers identify a neural mechanism that underlies some functional capacity; a strategy that requires one to descend at least one level (Piccinini and Craver, 2011). The prospects for success in this endeavor will improve if the range of environmental contexts within which one's mechanistic hypothesis must be tested is relatively narrow. As the range of contexts expands, the number of processes that might impinge on a hypothesized mechanism will increase, making the task of theory confirmation (or model validation) more difficult. In order to make neuro-functional investigation tractable, therefore, neuroscientists often think in terms of context-sensitive functional hypotheses, in which the number of environmental variables relevant to understanding a neural mechanism is minimized. Jacqueline Sullivan has made a similar claim and supported it with detailed case studies. Her formulation reads: "...given Craver's own commitment to the idea that mechanisms have specific components and temporal and spatial organizations, and given that such features of mechanisms are sensitive to what set that mechanism into motion, i.e., the events that preceded it, then it is entirely possible, indeed quite likely, that different experimental protocols will yield different mechanisms for potentially different phenomena" (Sullivan, 2009, p. 527).[2]

[1] My usage is intended to reflect the meaning of the term "context" as it is used in associative learning theory, and especially in classical conditioning.

[2] Sullivan exploits this feature of mechanistically oriented neuroscience in order to mount an argument against Craver's 2007 argument for the unity of the neurosciences. While I cannot take up that debate here, it should be mentioned that the solution I propose to the problem

Although context-sensitive neuro-functional hypotheses are legitimate and important aspects of neuroscientific investigation, I intend to show that, under some conditions, context-sensitivity systematically prevents such hypotheses from explaining neural structure. The argument for that claim is presented in the third section. In the remainder of this section I introduce and define the concept of task-bound functional analysis, a scientific strategy built around the concept of 'context-sensitive functional hypotheses.' Then, I distinguish two logically distinct forms of explanation in which task-bound functional analysis plays a central role. Finally, I highlight one of these strategies, and provide an example of how it works.

2.2 Defining TBFA

Task-bound functional hypotheses are a special case of context-sensitive hypotheses. In particular, they are hypotheses about the functional role of a brain area in which the kind of function posited has the property of being task-bound, which is defined as follows.

A function F is task-bound just in case (i) its description makes essential reference to a behavioral task invoked by an experimental paradigm in which environmental properties $P_1 \ldots P_n$ are manipulated, and (ii) it is ascribed to brain areas that are anatomically individuated.

Each component of this definition deserves clarification. In the first component, variables $P_1 \ldots P_n$ serve as partial identity criteria for an experimental paradigm. (I say partial because other factors may also be relevant to concerns about individuation, such as the content of the hypothesis under scrutiny.) The values of variables $P_1 \ldots P_n$ cannot be included in the definition of task-boundedness without sacrificing generality. This is because the criteria that determine the paradigm to which a particular experiment belongs are local to the empirical domain in question. Sullivan defines the term experimental paradigm as: "a standard method or procedure for producing an effect of a certain type" (Sullivan, 2009, p. 514). The exact set of environmental properties to be manipulated will depend on the nature of that intended effect. Typically these properties will include the kind of stimuli used and their temporal presentation, the shape and spatial layout of objects in close proximity to the organism, the

of task-bound functional analysis in Section 4 might be helpful in defending a version of the unity thesis. Thanks to an anonymous reviewer bringing this issue to my attention.

ambient lighting and acoustic conditions, and many other factors.[3] The crucial thing to note about the first component of the definition is that, among the identity criteria for a given experimental paradigm will be a set of environmental properties, the manipulation of which are required by the behavioral task invoked by that paradigm. The second component of the definition rules out analytic functional designations. Consider a claim of the following form: area x, new evidence suggests, does not realize function F1; in fact, it realizes function F2. If realizing function F1 is both necessary and sufficient for being counted as a token of type 'area x,' then this kind of claim is a matter of definition rather than empirical discovery. I am only interested in the empirical truth of task-bound functional hypotheses, so analytic statements must be set aside.

Given the definition of task-bound functional hypotheses, task-bound functional analysis, or TBFA, may be defined as a scientific strategy that attempts to taxonomize and explain the workings of the central nervous system (in any species) via the application of task-bound functional hypotheses.

2.3 The Direction of Explanation

I claim that TBFA is explanatorily limited. But the limitation I intend to identify only manifests itself when TBFA is deployed in the service of potential explanations with a particular logical form. In neuroscience, as in other fields, functional claims may be invoked either as explanans or explanandum. When invoked as explanandum, (probably the more widely discussed strategy in philosophy of biology) the existence of some function is already known, and then a structure is discovered that realizes the function, explaining how one particular system-type manages to achieve it. An example of this strategy in neuroscience is the explanation of photoreception in terms of the isomerisation of retinal and its subsequent liberation from opsin proteins in the rod and cone cells of the retina. It had long been known that the retina functions to detect and transduce light, but the structural realization of this function remained a mystery until the discovery that only one of two possible conformations of retinal binds to opsin

[3] In the same paper, Sullivan provides a thorough analysis of the components of experimental paradigms in the neuroscience of learning and memory. On her view, an experimental paradigm includes production procedures, which specify the experimental setup, measurement procedures, which specify relevant behavioral responses, and detection procedures, which indicate how empirical consequences can be attributed to dependent variables cast at the cognitive (as opposed to neurobiological) level. The environmental properties referred to in the definition of task-boundedness may fall under any of these three headings.

proteins (Wolf, 2001).[4] In this case, structural properties of retinal explain how the initial stage of transduction is realized. Let us refer to explanations of this kind, in which the discovery of structural properties explains how a system-type achieves some function, as functionally-oriented explanations. TBFA often makes possible the discovery of functionally-oriented explanations, and in this capacity, plays a crucial role in neuroscientific practice.[5] Much of the literature on mechanistic explanation illustrates exactly how this role has been, and ought to be played. (See, for example Craver (2007)).

In other kinds of explanation, the logical positions occupied by functional claims and structural claims are reversed, such that the discovery of a function explains why structural properties are present. For example, it was long known that some neural projections are covered with a layer of fatty tissue, known as myelin. But it was not until 1949 that Huxley and Stämpfli gathered the electrophysiological evidence required to demonstrate that sheaths of myelin, interrupted by the nodes of Ranvier, serve to create local circuits that propagate short electrotonic potentials down the length of the axon, making the transmission of current faster than it would be if the only mechanism of propagation were the chain reaction of action potentials, the speed of which is capped by the refractory period of voltage-gated sodium channels (Huxley and Stämpeli, 1949). This discovery explained not only why myelin sheaths are interrupted by the nodes of Ranvier (to create miniature circuits at internode intervals), but also why those sheaths have the material properties they do (in order to block current from escaping.) In this example, functional properties explain the presence of structural properties.

Non-scientific explanations of human artifacts often proceed in a similar logical fashion. Pinker (2009) relates a vignette in which a customer, browsing at an antique shop, stumbles upon a perplexing device composed of a stand, a metal arm bolted perpendicularly to the stand with a horizontally-oriented

[4]The history of this discovery is complex, but George Wald received the Nobel Prize for identifying the key molecular pathway that makes use of retinal in 1964.

[5]My use of the word 'structure' needs clarification. In logic and mathematics, the term often refers to an abstract set of relational properties. In this sense, propositions may have a logical structure that is independent of the meaning of non-logical terms, and a graph may have a topological structure that is independent of the identity of the individual nodes and the underlying geometrical space. In neuroscience, the concept of structure is to be contrasted with the concept of function. In this sense of the term, structural properties have to do with the shape and spatial organization of particular objects. At the scale of observable properties, usage of the term 'structural' in neuroscience is often synonymous with 'anatomical.' At the scale of unobservables, there exist non-anatomical structural properties, such as the conformation of a macromolecule.

ring on its end, a sharp x-shaped blade capable of pivoting vertically through the ring, and a lever to control the motion of the blade. The customer cannot understand why anyone would assemble such a bizarre collection of parts, until the shopkeeper explains that the device is an old-fashioned olive pitter. The ring holds an olive, and the lever drives the x-shaped blade through it vertically, pushing the pit through the flesh on the bottom of the olive and into a container below. Given that functional insight, the structural properties of the device suddenly make sense. Moreover, the sense of understanding achieved is not merely subjective. A person who understands the function of the olive pitter would almost immediately be in position to know what sorts of interventions would affect the performance of the device, and what sorts would not. Since the facts about whether an olive has been pitted are objective, information about the functional effects of potential interventions on the device constitutes a form of objective knowledge (as opposed to mere belief, or the subjective feeling of understanding that is sometimes elicited by the acquisition of knowledge). Let us refer to explanations of this form, in which the discovery of function explains the presence of structural properties, as structurally-oriented explanations.

In this paper, I restrict myself to the claim that TBFA faces limitations when used to generate structurally-oriented explanations; I prefer to set aside discussion of the potential limitations of TBFA when deployed in the service of functionally-oriented explanations.[6] In order to forestall confusion, I should emphasize that both 'functionally-oriented' and 'structurally-oriented' explanations are meant to denote forms of explanation that, one way or the other, include functional claims.

2.4 An Example of TBFA

With the above qualification in mind, it will be helpful to consider a well-known example of TBFA that has generated an impressive amount of scientific

[6]There is a growing philosophical and scientific literature on the forms of inference that are warranted by brain imaging data. One of the issues in this literature concerns the allegedly poor justification for what is known as "reverse inference." Reverse inferences are those that begin with the observation that brain area X is differentially activated in some behavioral task B. It is then noted that previously established results support the thesis that brain area X is involved in a particular cognitive process Y. From these two premises, it is concluded that behavioral task B engages cognitive process Y. The problems with this form of inference are intimately related to the limitations of TBFA when deployed in functionally-oriented explanations. However, philosophical analysis of this issue requires considerable discussion of brain-imaging technology, and would lead us too far afield. See Poldrack (2006) for an early criticism of reverse inference, and Hanson and Bunzl (2010) for an overview of the contemporary debate.

controversy.

The hippocampus is an anatomically defined structure in the medial temporal lobe of mammals and birds. It is widely agreed that the hippocampus serves a memory-related function, but consensus breaks down where more specific hypotheses are requested. As early as 1984, Nestor Schmajuk, a specialist in theoretical hippocampus research, was able to provide a list of 20 distinct theories of hippocampal function, and more have been added since (Morris, 2007; Schmajuk, 1984). Proposals include recording of spatial map coordinates, transitive inference, behavioral inhibition, and consolidating autobiographical memory (Per Andersen, 2007).

Perhaps what is most striking about the list of proposed functions is that it is far from obvious which properties, if any, all items on the list have in common. One reason for the diversity of proposed functions is, no doubt, the diversity of species upon which experiments are conducted: experimental paradigms designed for rats hardly resemble those designed for humans. Nevertheless, on the basis of widespread anatomical and physiological conservation, it is believed that the function of the hippocampus remains constant across species. Indeed, the structural similarities have prompted researchers to claim that a hypothetical surgery, replacing a human hippocampus with that of a macaque, could be entirely successful (Morris, 2007). This impressive degree of evolutionary conservation, combined with the fact that the organization of hippocampal substructures is characterized by an intricate form of layering that is unique within the mammalian brain, provide some reason to think there exists a function, at some level of description, that the hippocampus is expressly designed to perform. It is therefore puzzling, prima facie, that the existing hypotheses of hippocampal function are so wildly divergent. Moreover, it is puzzling that no unifying functional theories have emerged that show how the various task-bound theories are related to one another at a higher level of description.

By stepping back from the details of particular experiments, and focusing instead on shared methodology, it is not difficult to see how the application of TBFA makes the multiplicity of hippocampal theories more or less inevitable. The spatial coordinate mapping hypothesis is an attempt to explain behavioral phenomena observed in Morris water maze paradigms, in which a properly functioning hippocampus is required for remembering the location of a stable platform hidden below the surface of opaque water. The hypothesis that hippocampal function has to do with transitive inference is an attempt to explain behavioral phenomena in an odor-based paired association task, in which associ-

ations between bins of food cannot be consolidated over time without an intact hippocampus (Gilbert and Kesner, 2003). The hypothesis that hippocampal function has to do with behavioral inhibition is an attempt to explain the fact that rats with lesions to dentate gyrus cells fail to become still in the presence of what is normally perceived as danger (Takahashi, 1995).

The list of task-bound functional hypotheses could be extended, but these few examples suffice to illustrate the feature of TBFA that presently concerns us. It is this: each functional hypothesis in the literature is tailor-made to account for phenomena generated by one family of behavioral experimental paradigms. As the following section demonstrates, this kind of tailoring has the unfortunate consequence of desensitizing functional hypotheses to the data collected in paradigms that differ substantially from the one upon which a particular hypothesis is based.

3 Task-bound Functions Do Not Explain Structural Properties

3.1 Unipotent Structures

The first part of the argument is simple. Consider the strong assumption that each neuroanatomical structure—under a given description—realizes only one genuine function. Bastardizing terminology from developmental biology, I'll refer to this as the unipotent case.

If a neuroanatomical structure were unipotent, then the diversity of functional hypotheses proposed to explain the presence of structural properties would contradict one another. Since truth does not accommodate contradiction, and explanation requires truth, no explanation will be forthcoming until the contradictions between proposals can be resolved.[7] Moreover, in order to resolve those contradictions under the assumption of unipotency, any given anatomical area would have to be describable by one, uniquely true task-bound functional hypothesis. But no task-bound functional hypothesis is uniquely true. Since task-bound functions refer essentially to the environmental properties manipulated in an experimental paradigm, each task-bound hypothesis describes only

[7]The thesis that explanation entails truth was made explicit by Carl Hempel in his law-based account of scientific explanation Hempel and Oppenheim (1948). Despite the fact that Hempel's account is no longer thought to be as universally applicable as he had intended, this particular thesis remains popular with most accounts of scientific explanation. (Notable exceptions include those in Van Fraassen (1980), Achinstein (1985), and Bokulich (2011)).

a narrow range of the spectrum of functional contributions associated with an anatomical area. Given unipotency, then, task-bound functional analysis always yields contradiction, and for that reason, fails to explain the presence of structural properties.

3.2 Pluripotent Structures

The unipotency assumption is controversial.[8] It will be objected that neuroscientists do not always take incompatible functional hypotheses about the role of an anatomical structure to be logically contradictory. Sometimes, neuroscientists circumvent the incompatibility by indexing functional hypotheses to different environmental contexts. In light of this observation, it is necessary to consider the status of TBFA when applied to pluripotent neuroanatomical structures, where pluripotent means that the structure may serve multiple genuine biological functions over time. If a structure is pluripotent, there need not be any contradiction between what might otherwise appear to be competing functional hypotheses. Which particular function is realized at a time will depend on the occurrent behavioral goals of the organism, as well as the particular kind of environment the organism happens to be in. Howard Eichenbaum, a leader in hippocampus research, often expresses this ecumenical view: "... hippocampal sequence representation underlies a range of memory performance capacities, including episodic recall, cognitive mapping, sequence prediction, and inferential memory expression" (Fortin, Agster, and Eichenbaum 2002).

However, the possibility of multiple diachronically realized functions does not suffice to redeem the explanatory status of TBFA. To see this, recall the definition of task-bound functions: a function F is task-bound just in case its description essentially refers to a behavioral task in an experimental paradigm in which environmental properties $P_1 \ldots P_n$ are manipulated. Consider now what happens if the functional activity of the structure in question is investigated within a related experimental paradigm in which the set of manipulated environmental properties are different: $P_1 \ldots P_n$. Since task-bound functions are defined relative to experimental paradigms, a change in paradigm entails a change in the identity of the task-bound function. This has the consequence that new task-bound functions are all too easy to generate. Since there is no principled upper limit on the kinds of behavioral paradigms that might be invented,

[8] Some neuroscientists do take the unipotent view. With regard to hippocampal function, one of the best arguments for unipotency can be found in O'Keefe (1999), where the place-cell theory is advocated as the uniquely correct functional theory.

the list of purported task-bound functions for any given anatomical structure is unbounded: creative variations in the task description yield increasingly many new functions.

The possibility of an arbitrarily long list of functional ascriptions is the source of the explanatory weakness inherent in TBFA. There are two reasons for this. To see the first, assume that functions are defined relative to a containing system, as suggested by Cummins (1975). A containing system is a system that exhibits a complex capacity, relative to which the functions of its parts are defined.[9] (For example, the containing system for the pumping action of the heart is the circulatory system.) If we hold the containing system fixed, and if we are realists about functional roles, it should always be possible to construct a true (or approximately true) claim specifying the functional profile for a given neuroanatomical area. By functional profile, I mean a complete list of functional capacities. Faced with the specter of an unbounded list, however, such a construction is impossible. Any finite candidate list could be falsified by adding just one more functional ascription.[10]

Note that TBFA faces this problem even if neuroscientists have no empirical motivation to extend the list of hypotheses beyond its current length. For suppose that TBFA were the only method of discovering structurally-oriented explanations. Then, there could be no grounds for claiming of any list of task-bound functions that it is complete, even if empirical adequacy had been achieved.[11] Any functional contribution to behaviors yet to be taken into account would demonstrate the incompleteness of the list.

The second reason that the possibility of an arbitrarily long list of functions entails that TBFA is explanatorily limited is that there is a particular kind of understanding associated with the simplicity of functional insights that TBFA cannot deliver. Consider again the discovery of the function of myelin

[9] I do not mean to suggest that Cummins' theory of functional analysis is the appropriate one to use in this context. In fact, my account here is more compatible with what are sometimes called teleonomic functions, often associated with the work of Wright (1973). However, I do not think these two accounts of function are mutually exclusive, nor do I think that invoking the concept of a containing system commits one to either account.

[10] The canonical argument against the legitimacy of arbitrarily long disjunctions of realizers of functional kinds can be found in Fodor (1974).

[11] Since the theories in question concern the functionally relevant effects of a brain area, I assume that empirical adequacy is limited to those observations that incorporate neural data. Since neural data is only made available in an experimental setting, achieving empirical adequacy for a single theory implies very little about the vast majority of animal behavior. In order for a theory about one neural structure to yield predictions about behavioral data outside the confines of a particular experimental paradigm, commitment to a considerable body of additional theory is required.

sheaths. That discovery was powerful in part because the functional insight it represents is extraordinarily compact, and yet can account for the presence of at least two otherwise puzzling structural properties. An arbitrarily long list of functions would lack that conceptual compactness, which, although hard to quantify, seems essential to achieving significant explanatory depth.[12]

3.3 Objection I: Levels of Abstraction

One might suspect that the possibility of an unbounded list can be avoided by ascending to a more abstract level of description. If we characterize environmental properties abstractly, it is possible to derive a generalized functional hypothesis that stands in relation to specific task-bound functions as determinable stands to determinate. The generalized functional hypothesis will itself be more abstract than any given task-bound function it determines, and therefore invariant with respect to a larger class of changes in environmental properties. As the number of admissible changes increases, it becomes more difficult and less trivial to generate new functional hypotheses. For example, sequence prediction refers to a broad class of behavioral tasks used in learning and memory studies (Lisman and Redish, 2009). A wide variety of paradigms are counted as tests of sequence prediction processing capacities: a sequence of stimuli can be generated over spatial or temporal metrics; it can be either one or two-dimensional; stimuli can be visual, auditory, or olfactory, etc. We can define a neural function, F_S, as the function of carrying out sequence prediction processing. One specific instantiation of sequence prediction is a task in which the animal must visually detect non-random strings of stimuli buried within longer, random sequences Dunnett et al. (2012). By appealing to this more specific task, we can define another neural function, F_R, such that F_R is a determinate of the determinable F_S. Since F_S ranges over a broader class of environmental properties than F_R, it is surely less susceptible to the charge of explanatory weakness. So why not iterate the process of abstraction until a level is reached at which the number of possible functions is suitably limited?[13]

Although this iterative process of abstraction is, logically speaking, quite legitimate, it cannot redeem the explanatory status of TBFA. The amount of

[12] Unification-based theories of scientific explanation have tried to quantify this kind of conceptual compactness, with mixed results. See, for example, Kitcher (1981).

[13] Thanks to Colin Klein for suggesting that the determinate-determinable relation is applicable here. For a different argument about the limitations of abstraction, see Klein (2012). For empirical data on the wide variety of brain regions associated with memory storage, see Wager and Smith (2003).

explanatory strength to be gained by implementing such an abstraction process is severely limited by an additional constraint on structurally-oriented explanations: they must be capable of discriminating between distinct neural structures. If a purported explanation were true of multiple brain areas, each of which exemplified a unique form of structural organization, then, whatever else might be explained, it would not be the unique structural properties of one of those areas in particular. The need for this structural discrimination constraint can be clarified by analogy with the lock-and-key mechanisms of the immune system. Consider a purported explanation of the peculiar shape of an antibody that cites the fact that the antibody supports immune system function. Compare it to a purported explanation that cites the following fact instead: the antibody must be shaped as it is in order to ensure compatibility with the surface proteins on the antigen it serves to neutralize. All antibodies promote immune system function. In order to explain the structural properties of the antibody, a functional hypothesis must include specific information about the capacities of that particular molecule. The second of the two aforementioned hypotheses does include such information, and for that reason is clearly superior to the first. Similarly, all neuroanatomical areas participate in whatever function(s) are served by the nervous system as a whole, (i.e. survival) but that information is irrelevant to understanding the structural eccentricities of any particular area.

The explanatory limitations of very abstract functional hypotheses derive from the fact that they will be satisfied by many, and in some cases, all neuroanatomical structures in the brain. To return to our example: when it is claimed that the function of the hippocampus is sequence prediction, we say something which perhaps is true, but which fails to provide insight into its unique structural properties. Although originally proposed as a generalized theory of hippocampal function, sequence prediction is so general a concept that some researchers have ascribed a virtually identical function to the brain as a whole Friston (2010). and other proponents of the predictive coding theory of brain function argue that the most adequate general theory of brain function is the prediction of sequences of perceptual stimuli, given the memory of immediately past sequences. As we have already seen, however, if a functional theory applies so generally, it cannot shed light on the structural properties of the hippocampus in particular.

The lesson of this section is this: if we attempt to construct an abstract functional theory by generalizing over a large class of task-bound functions, the threat of an unbounded list is inevitable. In Section 4, I sketch an alternative approach to functional analysis that avoids both the explanatory limitations

associated with TBFA, and those of hypotheses that generalize over task-bound functions. Before moving on to that discussion, one additional objection should be considered.

3.4 Objection II: Containing Systems are Chosen by Convention

Another objection to the argument that TBFA cannot explain the presence of structural properties has to do with the distinction between proximal and distal effects. It can be framed as a reductio: the argument is absurd, one might think, because it threatens not only task-bound functions, but functional analysis generally. Compare the functional profile of a gene. If it is stipulated that the containing system for some activity of a gene is the cell nucleus, then the function of that gene will be specified in terms of proximal effects, such as the initiation of RNA transcription. If it is stipulated that the entire nervous system is the containing system, the function of the same gene will be specified in terms of distal effects, such as the regulation of neurogenesis. Some philosophers hold that containing systems are stipulated, rather than discovered. On this view, it is always possible to construct an arbitrarily long list of functions by stipulating additional containing systems. If such stipulation is legitimate, then this case seems parallel to the case of TBFA. If the two cases are parallel, then the argument in Section 3.2 ought to apply just as well to genetic explanation, and therefore gives us reason to think that no amount of functional information could provide insight into genetic structure. But if that is the case, the argument proves far too much, and should therefore be rejected.

This objection is mistaken for two reasons. First, in the gene case, the generation of new functions is constrained by the biological pathways that exist between gene and container. There may be many, but the set is finite.[14] In the case of TBFA, no such biological constraints exist. The number of unique behavioral tasks in which a neuroanatomical structure can be shown to participate is limited only by the creativity of the experimental designer. Second, and more importantly, in order to show that functional profiles generated by TBFA can be made arbitrarily long, it is not necessary to vary the containing system itself, as it is in the gene case. Rather, it is only necessary to vary the set of en-

[14]There are also important distinctions between biological pathways, such that some pathway-types permit explanatory relations to hold, while others do not. Woodward (2010) provides a good philosophical analysis of these type differences.

vironmental properties $P_1 \ldots P_n$ referenced in the definition of the experimental paradigm. So the variability in the TBFA case radically outstrips the variability in the gene case. Since the two cases are not parallel after all, the objection fails.

Arbitrarily long lists of function cannot explain the presence of structural properties. We have seen that TBFA makes impossible the identification of a principled upper limit on the list of functions associated with a given brain area. Any actual list generated by the application of TBFA will therefore have an arbitrary length (whether the magnitude of that length is great or not). Therefore, TBFA is incapable of explaining the presence of structural properties.

In the next section, an alternative concept of functional analysis in neuroscience is discussed, and some speculative remarks are made about the research programs that appeal to it. I argue that the form of functional analysis suggested by this alternative concept is not subject to the explanatory limitations of TBFA, and therefore deserves further investigation.

4 The Explanatory Depth of Intrinsic Brain Function

4.1 A New Concept of Neural Function

If the argument in Section 3 is correct, then the explanatory status of certain neuroscientific theories that depend on TBFA is threatened. Rather than trying to identify those theories, I propose a more positive, philosophical project: I propose to provide a rough sketch of a new kind of functional analysis that avoids the explanatory weakness of TBFA.

Given the vast amount of information yet to be learned about neural structure, this synthetic project must remain at least partially speculative. Nevertheless, it can be made systematic if it is framed in terms of a search for desiderata that a new, more explanatory concept of neural function ought to satisfy. The first of these desiderata is this: the new functional concept must denote a property capable of being realized in neural processes that span multiple behavioral tasks. This follows immediately from the argument in Section 3. As suggested above, in order to be explanatorily robust, neuro-functional hypotheses must not be susceptible to falsification by the replacement of one set of environmental properties with another. In order to accommodate alternative environmental

properties, a function must be realizable in processes that range across tasks.

4.2 Multiple Cognitive Domains

Recent work in functional neuro-anatomy suggests two additional desiderata. The first of these is inspired by the work of Michael Anderson, who has argued for a claim similar to the conclusion reached in Section 3, albeit on very different, mostly evolutionary grounds (Anderson, 2010, 2007)[15]. Although Anderson does not attempt to define a novel concept of neural function, he does make some suggestive remarks. "Insofar as our approach to discovering the specific functional role of a given brain area involves modeling its activity across different cognitive domains, then it makes little sense to try and characterize the contribution of the area using domain specific terms" (Anderson, 2007, p. 339).

Anderson is suggesting that the concept of neural function appropriate to capturing the functional role of anatomical areas must be one that is realizable in processes that span multiple *cognitive domains.*[16] This desideratum is stronger than the first, which only necessitated that our new concept of function be realizable in processes that span multiple behavioral tasks. To see that the former demand is stronger than the latter, one must only note the very small number of cognitive domains relative to the number of (possible) behavioral tasks. One standard taxonomy of cognitive domains includes just seven: attention, skill learning, semantic memory, language, episodic memory, and working memory (Cabeza and Nyberg, 2000). While finer-grained distinctions between cognitive domains are sometimes used, they are not nearly as fine-grained as the behavioral tasks invoked by standard experimental paradigms. Consequently, there will be a greater number of conditions under which one brain area is involved in two behavioral tasks than conditions under which that brain area is involved in two cognitive domains. Therefore, the requirement that a functional concept be realizable in processes that span domains is more stringent than the

[15] Anderson's "massive redeployment hypothesis" states that it is empirically likely that anatomically distinct brain areas developed in order to carry out one function, but, via exaptation, came to shoulder additional burdens throughout evolutionary history. Because the multiple etiological functions realized by one anatomical area can differ quite significantly from one another, the multiple reuse hypothesis is incompatible with the assumption that anatomically distinct brain areas are, generally speaking, dedicated to the realization of behavioral functions that are themselves defined with respect to a narrow spectrum of environmental contexts.

[16] Thanks to Jesse Prinz and Michael Anderson for helpful discussion about the way in which my thesis differs from Anderson's own.

requirement that it be realizable in processes that span behavioral tasks.

The motivation for imposing this condition is a body of empirical evidence from imaging meta-analyses that suggest that brain areas traditionally associated with one cognitive domain in fact participate in processes that span multiple domains (Cabeza and Nyberg, 2000; Wager et al., 2007). Such meta-analyses do not suggest that *every* brain area is multifaceted in this way, but they do show that multi-domain activity is common. If so, it would be useful to have a concept of neural function that can capture the general contribution of an area across cognitive domains. On this point, it is worth quoting Cabeza and Nyberg 2000 at length, from a passage in which they interpret the section of their results in which the data is modeled by anatomical region.

> Activations in parietal area 7, for example, were consistently found in studies of attention, space perception, imagery, working memory, episodic memory, and procedural memory. The most parsimonious account of this kind of activation is that they reflect cognitive processes that are tapped by tasks in different domains. However, most functional neuroimaging studies have preferred to interpret activations within their own domain... Area 7 activations, for instance, were usually attributed to attentional processes in attention studies, to perceptual processes in perception studies, to working memory processes in working memory studies, and so on. These domain-specific interpretations are useful because they allow researchers to refine hypotheses and plan new experiments... At the same time, it would be useful to systematically compare functional neuroimaging data in different cognitive domains and to develop general theories that account for the involvement of brain regions in a variety of cognitive tasks (Cabeza and Nyberg, 2000, pp. 31-32).

As Cabeza and Nyberg suggest, domain-specific functions are useful, but explanatorily limited. If we want a more general theory about the functional role of an anatomical area, we require a concept of neural function that is divorced from any essential connection to a particular cognitive domain.

4.3 Intrinsic Function

The third desideratum, which is also the strongest of the three, is this: our new concept of neural function must be capable of eschewing reference to behavior

altogether. This desideratum is motivated by a relatively new body of evidence suggesting that much neural activity is designed to achieve ends that are – in a sense that requires explication – intrinsic to the central nervous system. This formulation is inspired by the term 'intrinsic neural function,' which was recently coined by Marcus Raichle, and which he defines as "ongoing neural and metabolic activity not directly associated with the performance of a task" (Raichle, 2010). Raichle has marshaled evidence suggesting that at least a large proportion of neural function must be intrinsic in this sense. If he is right, then given the lack of association between this alleged class of neural functioning and the particularities of the behavior with which it is concurrent, we require a concept of neural function that can be articulated without mention of those particularities.

One line of evidence for Raichle's proposal comes from his own work on the default mode network — a network that is differentially activated whenever an organism is unencumbered by goal-oriented activity. For example, neuro-metabolic activity in the resting state (in humans) consumes approximately 20% of the body's total energy budget. Relative to resting state energy consumption, additional activity associated with momentary changes in brain activity is only about 5%, even during arousing perceptual and motor activity.[17] On the reasonable assumption that neuro-metabolic rates are correlated with functional operation, Raichle concludes that much of our normal brain function is driven by internal demands, rather than the transient, environmentally-modulated demands of whatever task an organism happens to be engaged in. Raichle's work focuses primarily on whole brain analysis, and contains little explicit discussion of the prospects for localization to anatomically defined areas. Nevertheless, Raichle's claims are based on calculations of glucose consumption, a process that is fundamental to the metabolic capacities of all neurons. Raichle's findings, therefore, are relevant to any theory of neural function, whether it concerns activities that are localized to small anatomical structures, or activities that are distributed across them. If, as this evidence suggests, the majority of localized neural functions are intrinsic, then we require a concept of neural function that is compatible with intrinsic functionality. I therefore suggest that the third desideratum be formulated as follows: the new concept must denote properties that are realizable in intrinsic processes.

It is important to note that, despite a superficial similarity, this condition

[17]The molecular evidence for this thesis is presented more fully in Raichle and Mintun (2006).

differs from the first—the one imposed by the result obtained in Section 3. That argument only shows that if we want structurally-oriented explanations, the neural functions we appeal to must be realizable in processes that range across behavioral tasks. In order to capture the contributions of intrinsic neural activity, the concept of neural function must be capable of eschewing reference to behavior altogether, which is a considerably stronger condition.

It should also be noted that the second and third desiderata constrain the concept of neural function in substantially different ways. Consider the (dubious) claim that the amygdala is a fear-processing center. Here, no reference to behavior is made, but the hypothesis still lacks explanatory power to the extent that the amygdala participates in processes that span other domains.[18] Because this case satisfies the second, but not the third desideratum, it serves to demonstrate the substantive difference between them.

4.4 Containing Systems Inside the Brain

Unfortunately, our three desiderata are difficult to satisfy without immediately running up against the limitation on abstraction discussed in Section 3.3. In order to provide non-trivial explanations of structural properties, (why are the inputs to the hippocampal formation mossy fibres, rather than pyramidal cells?) functional hypotheses must supply information that is both detailed and free from reference to particular tasks. The challenge posed by this requirement can be re-conceptualized as a tradeoff between the explanatory limitations of task-boundedness, on the one hand, and the explanatory limitations of ungrounded abstraction on the other.

Fortunately, it may be possible to avoid the tradeoff, at least in part, by reconsidering the kinds of containing system that are compatible with our desiderata. Since reference to behavior has been ruled out by the third desideratum, any containing system to which the new concept of neural function refers must be internal to the organism. In particular, the new concept of function will be defined with respect to containing systems inside the brain.[19]

[18] For a review of the fear hypothesis, see Rosen and Donley (2006).

[19] There is a growing literature on task ontology that can be viewed as an independent solution to the problem articulated in Section 3. The strategy behind constructing a task ontology is as follows: use imaging meta-analyses to find all the tasks that have elicited activation in one area. Then search for what these tasks have in common. (For an overview, see Poldrack et al. (2011)). My view is that this empirical strategy, in combination with computational modeling, is a very promising way of implementing the signal transformation approach I advocate here. As meta-analyses increase in sample size, we gain more insight into the functional profile of the area in question. Moreover, this method has lead quite directly to

This restriction forces us to think of neural function in terms of the proximal effects a brain area has on the circuit of which it forms a part. Since cognitive and behavioral vocabularies have been ruled out, we are forced to resort to a more abstract, mathematical level of description to capture those proximal effects. Rather than framing functional hypotheses about an anatomical area in terms of its contribution to behavior, or its capacity to realize a cognitive process, the new concept of neural function allows us to frame functional hypotheses in terms of signal transformation. Each anatomical hub in a large-scale brain circuit exhibits a unique set of structural properties that determines the kinds of transformation it is capable of carrying out. As a signal flows through the circuit, it will be transformed in a particular way at each anatomical hub. In principle, that transformation can be modeled mathematically. If so, functional ascriptions take on a logical form that is very different from those underlying most functional ascriptions in biology. Consider the following schema: the function of brain area A is to transform signal S in manner Φ. If transformation can be modeled by a mathematical function, then the functional role identified by this schema provides information that is both detailed and specific. If so, the tradeoff between task-boundedness and abstraction has been circumvented.

The two-variable formulation of the schema above is designed to make possible a relatively fine-grained form of functional analysis. Whether or not a brain region instantiates Φ will be determined by local microanatomy. But the individuation conditions on S may incorporate information about the sensory modality from which it originates, as well as information about the site at which the transformed signal is to be consumed. So even in cases in which two brain areas are indistinguishable with respect to local anatomy, they can nevertheless realize distinct functional roles.[20]

In the past decade, multiple new lines of research have adopted a signal transformation based view of functional analysis. What they have in common is their foundation in quantitative analysis of spiking neuron models. For example, Levy, Hocking, and Wu (2005) offer a theory of hippocampal function that

the kind of computational approach I advocate. For example, Penner-Wilger and Anderson (2008) use this method to introduce the concept of an "array of pointers" to show how one signal transformation is at work in the use of both finger and number representations.

[20] I believe that this proposal is compatible with the two-factor theory of neural semantics proposed in Eliasmith (2000). However, justification for that claim would require a detailed discussion of the relation between functional analysis and metaphysical semantics, which, being a controversial topic in its own right, cannot be included here. Sullivan Sullivan (2009) suggests that neuroscience would benefit from attempts to make the representational content of even low-level hypotheses more explicit. My schema could be taken as a sketch of a method for accomplishing that goal.

overcomes the limits of TBFA by appealing to the purely intrinsic concept they call the "associator of last resort." Briefly, the idea is that the hippocampus is designed to form associations between input patterns that are, statistically, so dissimilar from one another that other anatomical regions cannot "find" the associations.

Its ability to do this depends on the structural properties that define the network used to model the brain area, such as percentage connectivity and the degree of feedback inhibition from interneurons. These anatomical properties determine which functions (in the mathematical sense) a region is able to approximate.[21] In this functional hypothesis, the capacity that explains the presence of particular structural properties in the brain region is specified in terms of intrinsic statistical properties that remain constant over time and across environmental changes. Moreover, Levy's "associator of last resort" concept makes no reference to experimental tasks, particular cognitive domains, or particular behaviors, and thereby satisfies the three desiderata we have imposed on an explanatory concept of neural function.

I hope to have made two overarching points thus far. The first was that TBFA fails to explain the presence of structural properties. The second point was that the new concept of neural function suggests ways of overcoming that limitation. The philosophical upshot of these two points is simply that, when it comes to the brain, functional analysis will have to be carried out in abstract terms that are very different from those discussed in philosophy of biology. Localization of neural function is possible, but we will find the kinds of functions that can be localized quite surprising. Successful functional analysis will become integrated with what is currently known as "theoretical neuroscience" and will most likely draw on fields such as information theory, statistics, genetics, and perhaps physics, rather than directly from experimental paradigms in wet laboratories. Moreover, functions will be defined with respect to container systems within the brain itself.

These points return us to the claim made at the outset — that no brain area has a behavioral task as its principal function. Since task-bound functions cannot explain structural properties, they also cannot identify the unique con-

[21] For other brain areas, a similar perspective has been offered Rajesh Rao and colleagues. (See, for example Doya (2007)). This area of theoretical neuroscience has been growing exponentially over the past decade, and has provided a host of new concepts for neuroscientists to work with. This paper can be seen as an attempt to point out the why this mathematical literature should be considered relevant to concerns about functional analysis that have long occupied philosophers of biology.

tribution an anatomical area makes to the capacities of the brain as a whole. Since task-bound functions clearly fail to approximate that unique contribution, they ought not be used to label anatomical areas: the selection of any one label would be arbitrary. If we want our neuro-functional designations to be principled, they will have to satisfy the desiderata outlined above.

Acknowledgments

Thanks to Colin Klein, Michael Anderson, and Paul Humphreys for providing helpful comments on an earlier draft. Thanks also to Jesse Prinz and Matt Duncan for helpful discussion of the issues here discussed.

References

Achinstein, P. (1985). *The nature of explanation*. Oxford University Press.

Anderson, M. L. (2007). Massive redeployment, exaptation, and the functional integration of cognitive operations. *Synthese*, 159(3):329–345.

Anderson, M. L. (2010). Neural reuse: A fundamental organizational principle of the brain. *Behavioral and brain sciences*, 33(4):245–266.

Bokulich, A. (2011). How scientific models can explain. *Synthese*, 180(1):33–45.

Cabeza, R. and Nyberg, L. (2000). Imaging cognition ii: An empirical review of 275 pet and fmri studies. *Journal of cognitive neuroscience*, 12(1):1–47.

Craver, C. F. (2007). *Explaining the brain: Mechanisms and the mosaic unity of neuroscience*. Oxford University Press.

Cummins, R. (1975). Functional analysis. *J. Philos.*, 72(20):741–765.

Doya, K. (2007). *Bayesian brain: Probabilistic approaches to neural coding*. MIT press.

Dunnett, S. B., Fuller, A., Rosser, A. E., and Brooks, S. P. (2012). A novel extended sequence learning task (eslet) for rodents: validation and the effects of amphetamine, scopolamine and striatal lesions. *Brain research bulletin*, 88(2):237–250.

Eliasmith, C. (2000). *How neurons mean: A neurocomputational theory of representational content*. PhD thesis, Washington University.

Fodor, J. A. (1974). Special sciences (or: the disunity of science as a working hypothesis). *Synthese*, 28(2):97–115.

Friston, K. (2010). The free-energy principle: a unified brain theory? *Nature Reviews Neuroscience*, 11(2):127–138.

Gilbert, P. E. and Kesner, R. P. (2003). Localization of function within the dorsal hippocampus: the role of the ca3 subregion in paired-associate learning. *Behavioral neuroscience*, 117(6):1385.

Grodzinsky, Y. and Santi, A. (2008). The battle for broca's region. *Trends in cognitive sciences*, 12(12):474–480.

Hanson, S. J. and Bunzl, M. (2010). *Foundational issues in human brain mapping*. MIT Press.

Hempel, C. G. and Oppenheim, P. (1948). Studies in the logic of explanation. *Philosophy of science*, 15(2):135–175.

Huxley, A. and Stämpeli, R. (1949). Evidence for saltatory conduction in peripheral myelinated nerve fibres. *The Journal of physiology*, 108(3):315–339.

Kitcher, P. (1981). Explanatory unification. *Philosophy of science*, 48(4):507–531.

Klein, C. (2012). Cognitive ontology and region-versus network-oriented analyses. *Philosophy of Science*, 79(5):952–960.

Lisman, J. and Redish, A. D. (2009). Prediction, sequences and the hippocampus. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, 364(1521):1193–1201.

Morris, R. (2007). Theories of hippocampal function. In Andersen, P., Amaral, D., Bliss, T., and O'Keefe, J., editors, *The hippocampus book*. Oxford University Press.

Penner-Wilger, M. and Anderson, M. L. (2008). An alternative view of the relation between finger gnosis and math ability: Redeployment of finger representations for the representation of number. In *Proceedings of the 30th annual meeting of the Cognitive Science Society, Austin, TX*, pages 1647–52.

Per Andersen, David Amaral, T. B. J. O., editor (2007). *The hippocampus book*. Oxford University Press.

Petersen, S. E. and Fiez, J. A. (1993). The processing of single words studied with positron emission tomography. *Annual review of neuroscience*, 16(1):509–530.

Piccinini, G. and Craver, C. (2011). Integrating psychology and neuroscience: Functional analyses as mechanism sketches. *Synthese*, 183(3):283–311.

Poldrack, R. A. (2006). Can cognitive processes be inferred from neuroimaging data? *Trends in cognitive sciences*, 10(2):59–63.

Poldrack, R. A., Kittur, A., Kalar, D., Miller, E., Seppa, C., Gil, Y., Parker, D. S., Sabb, F. W., and Bilder, R. M. (2011). The cognitive atlas: toward a knowledge foundation for cognitive neuroscience. *Frontiers in neuroinformatics*, 5.

Raichle, M. E. (2010). Two views of brain function. *Trends in cognitive sciences*, 14(4):180–190.

Raichle, M. E. and Mintun, M. A. (2006). Brain work and brain imaging. *Annu. Rev. Neurosci.*, 29:449–476.

Rosen, J. B. and Donley, M. P. (2006). Animal studies of amygdala function in fear and uncertainty: relevance to human research. *Biological psychology*, 73(1):49–60.

Schmajuk, N. A. (1984). Psychological theories of hippocampal function. *Physiological psychology*, 12(3):166–183.

Sullivan, J. A. (2009). The multiplicity of experimental protocols: a challenge to reductionist and non-reductionist models of the unity of neuroscience. *Synthese*, 167(3):511–539.

Takahashi, L. K. (1995). Glucocorticoids, the hippocampus, and behavioral inhibition in the preweanling rat. *Journal of Neuroscience*, 15(9):6023–6034.

Van Fraassen, B. C. (1980). *The scientific image*. Oxford University Press.

Wager, T. D., Lindquist, M., and Kaplan, L. (2007). Meta-analysis of functional neuroimaging data: current and future directions. *Social cognitive and affective neuroscience*, 2(2):150–158.

Wager, T. D. and Smith, E. E. (2003). Neuroimaging studies of working memory. *Cognitive, Affective, & Behavioral Neuroscience*, 3(4):255–274.

Wolf, G. (2001). The discovery of the visual function of vitamin a. *The Journal of nutrition*, 131(6):1647–1650.

Woodward, J. (2010). Causation in biology: stability, specificity, and the choice of levels of explanation. *Biology & Philosophy*, 25(3):287–318.

Wright, L. (1973). Functions. *The Philosophical Review*, 82(2):139–168.