Function Follows Form: Biologically Guided Functional Decomposition of Memory Systems

David C. Noelle
University of California, Merced
5200 North Lake Road
Merced, California 95343
dnoelle@ucmerced.edu

Abstract

The field of design has offered the maxim that "form follows function," suggesting that the physical properties of an artifact should reflect our conception of its intended use. This advice can lead us astray, however, when looking to the brain for inspiration when developing intelligent systems. Too often, our preconceived notions concerning functional decomposition cause us to force strongly delineated functional roles on specific neural subsystems. To fully leverage the lessons of natural cognitive systems, we must remain open to functional architectures that might seem unnatural to us but are revealed in the form, or structure, of the brain. To illustrate this point, this report briefly reviews several nontraditional functional organizations of neural systems that have been proposed for working memory.

Introduction

The extremely flexible and general cognitive performance afforded by the human brain continues to make it the primary source of inspiration for the development of synthetic intelligent systems. As our understanding of neural mechanisms has advanced, a growing number of researchers have sought design insights from the functional organization of the brain. Modern theories concerning the structure of the human cognitive architecture have been shaped by both bottom-up empirical investigations into the dynamics of neural circuits and top-down functional constraints on the behaviors that humans produce. While both bottom-up and top-down considerations are needed to advance our understanding of how the brain manifests cognition, there is a danger in seeking extensive guidance from top-down information-processing analyses of cognitive processes when trying to sort out how neural subsystems are organized. Theoreticians regularly rely on intuitions shaped by experience with artificial information-processing systems to hypothesize particular functional decompositions of cognitive capabilities, suggesting that the problem of understanding the neural basis of cognition can be reduced to localizing each of the proposed components or modules in the brain and coming to understand the biological "implementation" of each. This "divide-and-conquer" approach is particularly appealing to those seeking to design highly complex artificial cognitive agents, and the constraints imposed by modern computing technologies (e.g., relatively limited parallel processing, impoverished sensors and actuators, lack of resources to support lengthy developmental approaches, etc.) frequently color the architectures that are proposed. Composing a cognitive architecture based on functional concerns, and then turning to the brain to understand how architectural components might be implemented, is a tempting approach with a long history (Marr 1982), but employing this strategy risks distracting us from evidence for functional architectures that differ from our preconceived notions.

In some cases, functional decompositions are suggested by behavioral data, carefully collected using the methods of experimental psychology. Differences in behavioral patterns under varying conditions are used to justify the existence of separate cognitive modules. Behavioral data is rarely conclusive in this regard, however, and there have been many computational demonstrations of how simple unified neural mechanisms can exhibit the same patterns of performance that have been used to argue for multiple distinct modules. In the domain of language, the apparent behavioral dissociation between rule-based processing and a memory for rule exceptions has been shown to arise naturally in neural systems that lack separate components for rules and exceptions (Plaut et al. 1996). The double dissociation logic of neuropsychological experiments, in which modularization is inferred from behavioral differences between patients with different focal brain lesions, has been critiqued on neurocomputational grounds (Plaut 1995). Behavioral data suggesting a hierarchical organization for skill knowledge, even in the case of routine tasks, has been explained by a neural model that lacks a separate "goal stack" and isolated representations of subskills (Botvinick and Plaut 2004). In these cases, and many others, behavioral data were originally taken as strong support for particular functional decompositions, only to be later shown to be consistent with neurocomputational architectures of a very different nature. These examples should act as warnings, suggesting caution when brain organization is viewed from the perspective of a prior assessment of functional demands.

In summary, we risk missing important insights into the cognitive architecture of the brain if we follow the famous dictate of the design disciplines: form follows function. We

can be easily misled if we start with a functional understanding of cognition and then search for our reasoned functional decomposition in the form, or structure, of the brain. To the degree that it is possible, we should allow function to follow form, allowing our understanding of the structure of specific neural systems to guide our understanding of the functional decomposition of human cognition. This can lead to novel insights at the architectural level, which may inform the broad design of intelligent systems.

This general point has been made many times before. It is at the heart of Braitenberg's "law of uphill analysis and downhill invention" (Braitenberg 1984), and it is a central theme of emergentist approaches to cognition, including much of connectionism. The goal of this brief report is to argue that allowing function to follow form when we theorize about brain organization can produce insights into human cognitive architecture that are practically useful for the design of artificial systems. This argument will be illustrated by some computational accounts of working memory performance. Some preliminary results are presented from a model of prefrontal cortex that learns to use something like "pointers," and this is followed by a discussion of the computational roles played by different brain areas in working memory.

Pointers In Prefrontal Cortex

Objects in the world often have parts, and it is intuitive to imagine that the brain's representations of objects would also allow for the encoding of the relationships between parts. There are many obvious examples, ranging from written words, which are composed of a sequence of letters, to simple rules, which are composed of a condition for rule application and an action to take. In computer science, there are a number of common strategies for encoding such compound objects. Simple compounds are typically encoded as structures that have a "slot-filler" organization, with a separate chunk of memory allocated for each "slot" and the bits in each chunk interpreted as an appropriate "filler" value. More complex compound data structures, such as dynamically sized lists, can be constructed by interpreting some fillers as "pointers" to other blocks of allocated memory. Unfortunately, it is unclear that these computer science data structures bear much resemblance to neural representations. There does not appear to be an addressable space of dynamically allocatable storage space in the brain. Instead, pools of neurons appear to be dedicated to a specific task by virtue of their connections within circuits. Still, a sort of slot-filler organization might be imagined, with specific neural assemblies representing static slots, and the pattern of neural firing within each assembly encoding a specific value. This view does not allow for the dynamic allocation of knowledge structures, but it does allow for the basic representation of compounds.

There is a general problem with this representational approach, however, which some researchers have called the "slot problem." This problem has several related aspects. First, dedicating pools of neurons to each structure component caps the maximum size of the full representation. Relatively unbounded objects, like a sequence of letters or the sequence of words that make up a sentence, cannot be represented without danger of running out of neural "slots" to fill. In general, this problem can be alleviated by dedicating many more pools of neurons than would ever be needed. But the second aspect of the slot problem cannot be solved in this way. If each slot involves an independent collection of neurons, then the learning of filler representations will have to be conducted separately for each slot. Learning about a filler in one slot will not transfer, at all, to the appearance of that filler in another slot. For example, learning to encode the letter "A" as the first letter of a word will not help the neural system learn to encode "A" as the second or third letter of a word, at all. Thus, if a given filler has never been experienced in a particular slot before, it is virtually impossible for the developing neural system to learn to properly encode that filler in that slot. Some limited success can be had by using distributed representations for fillers, allowing unexperienced fillers to be encoded based on their similarity to experienced fillers, but this solution still does not allow for any transfer from seeing a filler in one slot to its appearance in another. For example, learning the pronunciation of "A" as the first letter of a word would not transfer at all to learning how "A" should be pronounced as the third letter of a word. Finally, even if adequate experience is available to learn filler representations for every slot, this encoding scheme can be highly inefficient, as it requires that each slot include a sufficiently large number of neurons to reliably encode any possible filler. This duplication of resources can be expensive if there are many possible fillers (e.g., all possible subjects of a sentence being represented) and many slots.

Many of the proposed solutions to the slot problem have involved unrolling compositional structures in time. Recurrent neural networks can learn to process slot-filler pairs sequentially over time, integrating them into a fixed-size vector of neural firing rates (Pollack 1990; Sibley et al. 2008). These fixed-size representations can then be processed by other recurrent neural networks to extract components. Other proposed time-based solutions involve oscillating patterns of neural activity, with temporal synchrony used to indicate that a particular filler representation, encoded over one pool of neurons, is contained in a particular slot, encoded by a another pool of neurons. These time-based computational accounts might lead us to conjecture that the only way that the brain could appropriately encode compound knowledge structures is by utilizing some special temporal mechanism, traversing the structure slots in time either repeatedly as the structure is maintained or only when encoding and decoding a fixed-size neural representation of the structure.

Another option appears, however, when we look at the prefrontal cortex (PFC) of the brain. The PFC is thought to play an important role in working memory (Goldman-Rakic 1987). Neurons in this brain region have been found to actively maintain high firing rates in the absence of stimuli, encoding relevant bits of information during delay periods. Many different kinds of information appear to be actively maintained in the PFC, including spatial locations (Funahashi, Bruce, and Golman-Rakic 1989), recently viewed objects (Cohen et al. 1994; Miller and Des-

imone 1994), action rules (Wallis, Anderson, and Miller 2001), and even verbal information (Demb et al. 1995). Unusually dense recurrent connections in PFC are thought to support active maintenance of high firing rates through mutual excitation (Camperi and Wang 1998). Interestingly, there is anatomical evidence of a "slot-filler" structure in the PFC. Collections of recurrent excitatory connections appear to be limited to isolated stripe-like collections of neurons, producing pools of neurons whose firing rates can be independently maintained (Levitt et al. 1993; Pucak et al. 1996). Loop-like projections from the PFC through the basal ganglia, the striatum, and the thalamus are thought to provide a mechanism via which these stripe-like neural pools can be independently updated (Frank, Loughry, and O'Reilly 2001).

Given this "slot-filler" anatomical structure in PFC, without any apparent temporal binding method, it is natural to wonder how the PFC resolves the slot problem. One possible answer to this question arises from the observation that the neural "loops" from PFC through the basal ganglia are somewhat asymmetric, with neural stripes in more anterior parts of PFC affecting portions of the striatum that control the updating of more posterior stripes in PFC, but not vice versa. With this pattern of connectivity, the pattern of firing rates in more anterior stripes could encode the identity of specific more posterior stripes, rather than a "filler" pattern, directly. In this way, the anterior PFC stripes might be seen as encoding a kind of "pointer" to more posterior PFC stripes. Thus, the slot problem could be addressed by having a relatively small number of more posterior PFC stripes learn to encode each possible "filler" value, avoiding the need for the properties of "filler" values to be learned separately for every possible "slot" location. More anterior PFC stripes could learn to encode "slots" for structured knowledge, referencing the more posterior PFC stripe that currently contains the corresponding "filler" value. This kind of "indirection" could resolve the slot problem without recourse to a special temporal mechanism.

Building on previous models of PFC (Hazy, Frank, and O'Reilly 2006), we have begun to implement a computational cognitive neuroscience model of the interactions between anterior PFC stripes and more posterior PFC stripes during simple working memory tasks. Like the models that it extends, this model is implemented using the Leabra framework (O'Reilly and Munakata 2000), which uses firing-rate neurons and biologically plausible mechanisms for synaptic plasticity, including a reinforcement learning mechanism embodied by the brain's dopamine system. The model includes the observed asymmetric connections between more anterior PFC stripes and more posterior PFC stripes, but it is otherwise largely unconstrained in what information is encoded in the neural pools that make up each stripe. Standard neural learning mechanisms shape the stripe representations as the model is trained on a simple paired-associate working memory task where the same set of cues are always used, corresponding to "slots," but the cues are associated with different "filler" values on different trials. Preliminary results indicate that networks without the anatomically motivated asymmetry between anterior and posterior PFC stripes can generalize fairly well on this task, but generalization fails dramatically when specific "fillers" never appear in specific "slots" during training. This kind of generalization is successful, however, when more anterior PFC stripes are allowed to learn to act as "pointers" to more posterior PFC stripes. Thus, the slot problem is largely solved by introducing a bit of anatomical hierarchy.

Any computational architecture for synthetic cognitive agents will surely need methods for representing compound structured knowledge in a manner that also supports flexible learning and generalization. Based on previous explorations with connectionist models, it would be tempting to insist that any biologically inspired architecture should include a fundamental mechanism for unrolling structures in time. An examination of the functional anatomy of PFC suggests an alternative approach, however. Similarly, computational intuitions might stress the great utility of "pointers" in artificial systems, suggesting that these should be a fundamental element of any cognitive architecture. Initial results from our ongoing modeling efforts, however, suggest that pointer-like functionality might be learned, emerging from an interaction between experience and simple anatomical constraints rather than being hard-wired and fixed.

Decomposing Working Memory

Contemporary theories of human memory have frequently aligned specific brain systems with behavioral distinctions concerning kinds of memory performance. Some researchers would assign working memory to the PFC, "declarative" or "episodic" memory to the hippocampus in the medial temporal lobes, and "procedural" or "skill" memory to more posterior regions of neocortex, for example (Squire 1987). As we design cognitive architectures for artificial agents, it is tempting to partition memory into categories of this kind and assign separate architectural components, or modules, to each kind of memory. This strategy would then send us to particular brain regions to uncover biological insights into how each module is best implemented.

There are a growing number of reasons to be suspicious of such simple assignments of behaviorally-delineated memory types to specific brain regions, however. It is increasingly clear, for example, that PFC plays an important role in episodic memory encoding and retrieval. Conversely, the hippocampus can be central to performance on some working memory tasks. Examining the structure of the brain may very well suggest a different functional decomposition of memory systems.

There is unusually strong lateral inhibition in key regions in the hippocampus, such as the dentate gyrus, producing patterns of neural firing that are exceptionally sparse. Computational analyses of such sparse neural representations suggest that they would support poor generalization to novel situations, but would have other useful properties, such as the ability to be learned quickly without producing unwanted interference (McClelland, McNaughton, and O'Reilly 1995). Thus, the hippocampus might be particularly good at maintaining quickly formed, perhaps even one-shot, memory traces that are very specific in nature. General cortical association areas, in comparison, might be par-

ticularly good at maintaining general knowledge, acquired slowly over repeated experiences. PFC, with its particularly dense pattern of recurrent excitation, might be particularly good at actively maintaining a memory trace as a pattern of neural firing. This "tripartite organization" scheme — separating computational contributions between hippocampus, PFC, and other cortical areas — is central to many Leabra-based accounts of brain function (O'Reilly and Munakata 2000). With this sort of functional decomposition of memory systems, all of these brain regions might be simultaneously involved in a typical memory task. For example, a common working memory task, such as remembering a phone number, might be simultaneously supported by an active pattern of neural firing in the stripes of PFC, by a quickly generated and highly specific memory trace in hippocampus, and by the familiarity of certain number subsequences (e.g., "123") appearing in the phone number, as embedded in the synaptic connections of posterior cortex. Thus, under this view, no one brain region is responsible for working memory, per se (O'Reilly, Braver, and Cohen 1999).

If this functional decomposition of memory systems is at least somewhat accurate, it would be misleading to focus on specific brain regions when seeking biological inspiration for the design of, say, a working memory system, or a declarative memory system, or a procedural memory system, or the like, for an artificial agent. Indeed, this neurocomputational view might suggest a fundamentally different memory architecture for synthetic agents — one in which different modules cooperate across a variety of memory tasks.

Conclusion

This brief article argues that great care should be taken when looking to the brain for inspiration for the design of artificial cognitive architectures. Specifically, there is a danger in composing a functional decomposition of cognitive capabilities based only on computational or behavioral considerations, turning to the brain only to uncover insights into the implementation of pre-specified modules. Examining the gross structure — the form — of the brain can lead to novel insights into the general functional architecture of human cognition. In this paper, this point is illustrated in the context of pointer-like representations that are thought to emerge in prefrontal cortex and in the context of general memory system organization. These examples encourage us to allow function to follow form as we design biologically inspired cognitive architectures.

Acknowledgments

Ongoing work on indirection in prefrontal cortex is being conducted by Trent Kriete in the Computational Cognitive Neuroscience Laboratory at the University of California, Merced in collaboration with the author of this report, Randy O'Reilly, Jonathan Cohen, Todd Braver, Alex Petrov, and Michael Frank. Thanks are offered to Alexei Samsonovich for inviting this submission.

References

Botvinick, M., and Plaut, D. C. 2004. Doing without schema hierarchies: A recurrent connectionist approach to routine sequential action and its pathologies. *Psychological Review* 111:395–429.

Braitenberg, V. 1984. *Vehicles: Experiments in Synthetic Psychology*. Cambridge, Massachusetts: MIT Press.

Camperi, M., and Wang, X.-J. 1998. A model of visuospatial working memory in prefrontal cortex: Recurrent network and cellular bistability. *Journal of Computational Neuroscience* 5:383–405.

Cohen, J. D.; Forman, S. D.; Braver, T. S.; Casey, B. J.; Servan-Schreiber, D.; and Noll, D. C. 1994. Activation of prefrontal cortex in a nonspatial working memory task with functional MRI. *Human Brain Mapping* 1:293–304.

Demb, J. B.; Desmond, J. E.; Wagner, A. D.; Vaidya, C. J.; Glover, G. H.; and Gabrieli, J. D. E. 1995. Semantic encoding and retrieval in the left inferior prefrontal cortex: A functional MRI study of task difficulty and process specificity. *Journal of Neuroscience* 15:5870–5878.

Frank, M. J.; Loughry, B.; and O'Reilly, R. C. 2001. Interactions between the frontal cortex and basal ganglia in working memory: A computational model. *Cognitive, Affective, and Behavioral Neuroscience* 1:137–160.

Funahashi, S.; Bruce, C. J.; and Golman-Rakic, P. S. 1989. Mnemonic coding of visual space in the monkey's dorsolateral prefrontal cortex. *Journal of Neurophysiology* 61:331–349.

Goldman-Rakic, P. S. 1987. Circuitry of the prefrontal cortex and the regulation of behavior by representational knowledge. In Plum, F., and Mountcastle, V., eds., *Handbook of Physiology*. Bethesda, MD: American Psysiological Society. 373–417.

Hazy, T. E.; Frank, M. J.; and O'Reilly, R. C. 2006. Banishing the homunculus: Making working memory work. *Neuroscience* 139:105–118.

Levitt, J. B.; Lewis, D. A.; Yoshioka, T.; and Lund, J. S. 1993. Topography of pyramidal neuron intrinsic connections in macaque monkey prefrontal cortex (areas 9 46). *Journal of Comparative Neurology* 338:360–376.

Marr, D. 1982. *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*. San Francisco: W. H. Freeman.

McClelland, J. L.; McNaughton, B. L.; and O'Reilly, R. C. 1995. Why there are complementary learning systems in the hippocampus and neocortex: Insights from the successes and failures of connectionist models of learning and memory. *Psychological Review* 102:419–457.

Miller, E. K., and Desimone, R. 1994. Parallel neuronal mechanisms for short-term memory. *Science* 263:520–522.

O'Reilly, R. C., and Munakata, Y. 2000. *Computational Explorations in Cognitive Neuroscience: Understanding the Mind by Simulating the Brain*. Cambridge, Massachusetts: MIT Press.

O'Reilly, R. C.; Braver, T. S.; and Cohen, J. D. 1999. A biologically based computational model of working memory. In Miyake, A., and Shah, P., eds., *Models of Working Memory: Mechanisms of Active Maintenance and Executive Control*. Cambridge: Combridge University Press. chapter 11, 375–411.

Plaut, D. C.; McClelland, J. L.; Seidenberg, M. S.; and Patterson, K. 1996. Understanding normal and impaired word reading: Computational principles in quasi-regular domains. *Psychological Review* 103(1):56–115.

Plaut, D. C. 1995. Double dissociation without modularity: Evidence from connectionist neuropsychology. *Journal of Clinical and Experimental Neuropsychology* 17:291–321.

Pollack, J. B. 1990. Recursive distributed representations. *Artificial Intelligence* 46(1–2):77–105.

Pucak, M. L.; Levitt, J. B.; Lund, J. S.; and Lewis, D. A. 1996. Patterns of intrinsic and associational circuitry in monkey prefrontal cortex. *Jounral of Comparative Neurology* 376:614–630.

Sibley, D. E.; Kello, C. T.; Plaut, D. C.; and Elman, J. L. 2008. Large-scale modeling of wordform learning and representation. *Cognitive Science*. in press.

Squire, L. R. 1987. *Memory and Brain*. New York: Oxford University Press.

Wallis, J. D.; Anderson, K. C.; and Miller, E. K. 2001. Single neurons in prefrontal cortex encode abstract rules. *Nature* 411:953–956.