Evolutionary Boxology

Working notes from Gerald Loeb, University of Southern California, 9/2011 – 8/2012

When students of sensorimotor behavior venture into theorizing about their system, they generally employ a method known somewhat facetiously as “boxology” – the construction and labeling of a set of interconnected “black boxes” to which are ascribed certain functions. Such schematic representations are commonly used in engineering when designing new systems for which a theory of operation can be imposed by the inventor. Discrete boxes also correspond to the anatomically discrete structures of the nervous system, so that boxology is nearly unavoidable in our efforts to understand the functions of the nervous system. However, as David Marr (1982) famously pointed out, such boxology can be a trap when attempting to discover how a functioning system actually works because it tends to focus our attention on familiar or convenient computational approaches to a problem and to lead us away from the question of the theory of computation actually employed by the organism. Theoretical sensorimotor control now consists of an embarrassing proliferation of arrangements of boxes that all account reasonably well for substantial sets of experimental observations. Each of these constitutes an embodiment of one theory of computation. When new observations cannot be accounted for by one such embodiment, the nervous system offers a sufficient number of anatomically distinct structures whose functions remain unclear; these can then be reassigned within the model to preserve its plausibility and completeness, at least conceptually. But suppose the theoretician had started with a different embodiment, a different arrangement of boxes. Then the unaccounted observations would be different and the new boxology would have a different connectivity and a different functional label. In time, there will be a proliferation of theories of computation, each able to account for normal behavior.

An engineer would be pleased to have developed so many useful inventions and would select among them by some simple strategy such as comparing how well they optimize some cost function of relevance to a given application (e.g. speed, accuracy, energy consumption, storage requirements or combinations thereof). A biologist attempting to discover the one true theory of computation of a particular organism would be dismayed. Removing one of the boxes and looking at functional deficits has provided most of the guesses that underlie the current boxology, but the reciprocal connections and modulatory functions among neural structures limit the ability to assign credit or blame to specific structures. The biologist might resort to probing each of the boxes to see if the signal processing that can be observed in that box is consistent with its assigned functional label. Unfortunately, the biological boxes are each very complex, their signals are very noisy, and the whole experimental method is geared only to assessing the relative probabilities of various different but known hypotheses. The data can never exclude the possibility that the actual function of the structure being probed corresponds to a different and unknown or unconsidered theory of computation.

Fortunately, biology provides a constraint on boxology that might winnow the proliferation of theories of computation more effectively than reductionist probing and more appropriately than optimizing an arbitrary cost function. Species and their nervous systems evolved in an orderly manner, along with their behavioral capabilities (Loeb, 1999 [1]). Earlier brains are simpler, with fewer anatomical structures to
be included in the boxology and with more limited motor repertoires, albeit with impressive capabilities. Interestingly, brains supporting simpler and more complex behavioral repertoires actually have similar architectures. Generally, a more complex repertoire is achieved by growth in the relative size of certain brain regions or even the development of new neural structures, but either must function in the context of a generally preserved neural architecture. New structures and capabilities must improve specialized performance by the organism without compromising the effectivity of earlier structures. As the new structures develop phylogenetically, they may (or may not) continue to rely on the older structures for some functionality but they must suppress the older structures’ ability to operate autonomously. By looking at the correlation of behavioral capabilities and anatomical development, we can gain insights into the biological boxology.

In the following, we consider what changes and what does not as organisms evolve from loosely connected sets of individual ganglia controlling discrete and local functions (invertebrates) to precisely directed movements according to specialized senses (e.g. cold-blooded vertebrates) to the ability to learn, refine and select from a repertoire of optional behaviors (e.g. mammals and ultimately primates). The boxology in Figure 1 represents the anatomical substrates that appear to account for much (admittedly not all) of the evolving behavioral repertoire. We will attempt to explain why these structures evolved, for what behavioral advances they account, and what function each performs.
The nervous systems of **invertebrates** consist largely of anatomically local and functionally autonomous ganglia for body parts that are coupled together in segmented chains. Coordinated motor behaviors emerge from mechanical coupling between the body segments and fixed reciprocal connections between the ganglia. Special senses such as antennae and eyes are processed by a rudimentary brain to trigger one of a few stereotyped behaviors such as prey-catching or escape. Aside from homeostatic mechanisms to maintain stable behavior, there is little or no ability to learn new behaviors or refine old ones.

In **earlier vertebrates**, the various ganglia have become fused into an elongated and continuous spinal cord, but the projections to and from the musculoskeletal plant are still relatively localized. The specialized senses are processed primarily by the tectum, which translates them into accurately directed behaviors to acquire targets. Acquisitive behaviors such as gaze shifts, tongue spearing and lunging require coordination of the whole musculoskeletal plant. Thus the local spinal circuits must be augmented by a similar but more widespread coordination network, particularly in terrestrial animals that must stabilize the body against gravity and aim special sense organs such as eyes and ears that are mounted at the end of a highly articulated neck and trunk. Brainstem structures such as the vestibular nuclei and reticular formation provide this coordination, but they must be carefully tuned to optimize accuracy for the particular behaviors and circumstances that turn out to be most important to the organism. Furthermore, they must be retuned as the animal grows (ontogeny) and they must cope with the changes in body form and habitat that are the **sine qua non** of evolution (phylogeny). This requires an adaptive controller that can integrate information from all exteroceptive and proprioceptive sensors plus internal states such as spinal interneuron and motoneuron activity. The cerebellum provides this new functionality. Nevertheless, the behavioral repertoire itself is still largely preprogrammed by genetically specified central pattern generators. As the primitive cerebral cortices of reptiles gradually enlarge, the first indications of learned motor behaviors emerge, perhaps represented by the mimicky associated with several unrelated species of birds.

In **vertebrates** that evolved flexible behavioral repertoires, all of the structures subserving the standard behavioral repertoire are present and functional, but their function is increasingly under the control of the cerebral cortex. Exteroceptive sensory information that used to trigger automatic motor behaviors via the tectum is routed also through cerebral cortical areas that learn to evaluate and classify according to context and prior experience, producing ever more abstract representations of the sensory information. Motor behavior constitutes the highest abstraction of sensory information – what to actually do in a given situation. The individual elements of even a complex, learned motor sequence are likely to be similar to patterns of motor coordination that other parts of the nervous system are already well-equipped to perform with grace, speed and efficiency. What the cerebral cortex adds is the ability to decide when to begin and how to select and sequence those elements. It also must learn to control directly the recently evolved musculoskeletal specializations such as opposable fingers and an articulatory vocal tract, for which no lower motor control centers ever evolved.

In order to assert control over sensorimotor behavior, the cortex must delay or modulate the responses of other motor centers in the tectum and spinal cord to ensure that multiple responses from different control centers do not conflict. This implies a pervasive inhibitory bias of the lower
centers that is then lifted if and when the cortex decides to go ahead with the sensorimotor behavior. That inhibitory bias tends to obscure the existence and potential importance of the lower circuits when they are probed in reduced or non-behaving preparations. But its existence accounts for experiments in which the cortex plans and controls virtual limb movements via decoded neural activity without generating overt muscle recruitment.

The ability to think about possible motor behaviors without actually emitting them would enable a form of learning that primates appear to have exploited most thoroughly. Neural networks have the wonderful property of eventually producing any rewarded output, but “eventually” constitutes a fundamental flaw in the real world. A neural network can only be trained by exposing it to input data and to the consequences of its outputs over large numbers of trials, which may be hazardous to the organism’s health. If the organism can store a snapshot of the input conditions for which it has no satisfactory output and then replay the scenarios offline (e.g. thinking or dreaming), then it can train its neural networks intensively at lower cost and risk, a huge evolutionary advantage. The ability to do this is cumulative because it depends on an increasingly rich and accurate repertoire of expected outcomes from various hypothetical actions.

The increasing tendency for primitive elements to be biased “off” in higher mammals and particularly primates helps to understand how this boxology generalizes beyond traditional sensorimotor behaviors to other cognitive, strategic and emotive behaviors of the brain. This is important, because the basic boxology presented here seems to apply across a huge range of these functionally distinct behaviors and the cortical and subcortical structures with which they are associated. From this perspective, the classical Freudian battle between id and superego becomes the battle between the limbic, reptilian brain and the prefrontal cerebral cortex that attempts to impose learned gating on innate behaviors that may be inappropriate under some but not all circumstances. The anatomical substrates and physiological processes are not fundamentally different from the ability of the visual cortex and prefrontal eye fields to countermand the much faster express saccades that both humans and lower vertebrates can make using the retinotectal pathway of the superior colliculus. Interestingly, humans exhibit a wide range of such inhibitory biasing across individuals, which at their extremes account for much of traditional psychopathology: impulsiveness, aggressiveness, exhibitionism, attention deficits on one end and perhaps autism, obsessive-compulsive disorders on the other.
The figure above expands on the mammalian scheme to include the remainder of the main extrapyramidal structures that are known to contribute to normal and pathological sensorimotor behavior. A summary putative function for each box is provided in italics under each anatomical name. The main refinement is a division of cerebral cortical function into “perceptual” and “proactive”, as opposed to the more common descriptions of “sensory”, “motor” and “associative”. As I.T. Diamond pointed out (“The subdivisions of neocortex: A proposal to revise the traditional view of sensory, motor, and association areas”, Prog. In Psychobiol. And Physiol. Psychol. 8:1-43, 1979), all cortical areas have similar neurobiology and types of input/output connections and probably employ the same computational and learning algorithm. All cortical areas probably contribute something both to perception (the development and selective recall of a repertoire of familiar entities based on prior experience) and to proactive behavior (the generation of learned actions that result in particular percepts based on prior experience with those entities). When we
present visual cortex with visual stimuli and ask it to make classifications while ignoring its ability to select saccadic gaze shifts, we force ourselves to treat it as perceptual rather than proactive cortex. When we decode motor cortical activity in terms of observable voluntary movements while ignoring its ability to learn which behaviors to generate, we force ourselves to treat it as proactive rather than perceptual cortex. By recognizing that both functions probably coexist in ALL cortical areas, we can start to understand the roles of three other subsystems that each connect to virtually ALL cortical areas:

- **Hippocampus:** As noted above, when a neural network is exposed to an unfamiliar input pattern, it will likely take many iterations of Hebbian learning gradually to classify the pattern as a variant of something already experienced or as a new assemblage of more primitive percepts. Some part of the brain has to know when such unresolved situations arise and must store and then reconstruct the raw data for offline thinking – “rumination”. For various reasons (to be elaborated elsewhere), it seems plausible for this necessary function to be performed by the hippocampus and its widespread and rapidly potentiable projections to cerebral cortex. This scratchpad storage function would be handy for performing the most studied functions supported by the hippocampus – maze-running and other memorizations and recalls of arbitrary sequences of stimuli.

- **Basal Ganglia:** If the proactive function of cortex is learned on the basis of reinforcement (operant conditioning), then some structure has to evaluate the rewards. If that structure also has a memory and learning function, then it is in a position to predict the reward value of any action. Such a vetting function on motor cortical output (i.e. actions) would account for the akinesia of Parkinson’s disease, essentially a blockade of the proactive function of cortex. More speculatively, we can consider the perceptual function of cortex to result in decisions to classify or not. Recognizing a complex situation as familiar can be construed as rewarding; being stymied may be quite threatening. Thus, the basal ganglia may be in an excellent position to inform the hippocampus when a situation bears remembering and rumination (hence the negative projection from basal ganglia to hippocampus in the figure denoting absence of the reward of recognition). The basal ganglia connect similarly to both “sensory” and “motor” cortical areas, which further emphasizes their computational similarities and proactive reward-seeking behavior.

- **Cerebellum:** The various actions that result from actions initiated by cortical areas (i.e. everything under voluntary control including both overt movements and shifts of attention) are mostly elaborated dynamically by lower subsystems (tectum, brainstem, spinal cord). Each lower subsystem is wired for rapid control of specific functions but it lacks the “big picture” of the consequences of its actions for function in the other subsystems. This problem can be appreciated at a simple mechanical level by Coriolis forces or intersegmental dynamics, particularly in limbed organisms. Any attempt to move one joint has wide-spread consequences for all other segments in the linkage, but those consequences depend strongly on the musculoskeletal posture and tone at each of those other joints. One way to solve this problem would be to include state feedback from the other subsystems as part of each subsystem’s inputs. In order to represent all possible conditions of the entire organism, however, the number of such integrative elements would be impossibly large for the available space and for the cortical areas to learn to control. The cerebellum is a giant matrix switch of state information that can mix any combination of lines but not all possible combinations simultaneously. Its job is to learn which combinations are useful for which circumstances and to invoke those combinations and provide them to the appropriate
subsystems when those circumstances arise. Thus it needs inputs that tell it what circumstance obtains (mostly from proactive decisions of cerebral cortex in mammals), state information from lower subsystems (their afferent input and efferent copy signals), and error signals to train its neural network to produce the appropriate outputs during each of those circumstances.

It is common to study and interpret these subsystems in the context of particular overt sensorimotor behaviors, but the functional roles described above can be generalized to all capabilities, including purely cognitive and emotive conditions as mentioned above. Even when viewed from the narrow perspective of sensorimotor behaviors, each one of these “extrapyramidal” subsystems performs functions that cannot be understood without an accurate representation of the functions of all of the other subsystems, including both the lower subsystems that elaborate the behaviors and the higher subsystems that decide which behaviors are required. A proposed boxology that includes all of those functions can then be used to generate experiments that test its own viability, by predicting which signals and connectivity should be present where and which should not. Note, however, that as in any hypothesis testing, the experimental data can only falsify the viability of the model structure. No matter how many consistent data are accrued, they cannot be used to assert that a given boxology is correct because those data may happen also to be consistent with another, unknown or unconsidered boxology. Unfortunately, as discussed in the box below, a certain amount of inconsistent data is almost inevitable, so cannot be taken immediately at face-value as a refutation of a given model.
A Digression and Caveat about Hypothesis Testing for Boxology: The relationship of experimental to theoretical biology (i.e. data to boxology) is much more tenuous than for experimental to theoretical physics. Biological data are much noisier and more likely to be erroneous or at least misleadingly interpreted than physical data. The noisiness is obviously inherent in the biological processes themselves, compounded by the variability in deployment and operation of those processes in a given population of organisms, and obscured by the extremely limited and often biased sampling of the internal states that is experimentally feasible (e.g. single unit recording). The erroneousness is more of a nasty little secret. Biological experiments are hard to do and they are usually done by individuals and small teams in support of their own theories, so bias is hard to exclude and harder to detect. The interpretation of results usually hinges on excluding alternative possibilities on the basis of a large body of prior knowledge about anatomy and physiology, but that body is never complete at the time the results are published and is itself subject to noise and error. The best scientists discuss all of these alternative possibilities and sources of uncertainty and error in their journal articles, but the articles are usually remembered and summarized in reviews on the basis of the most likely interpretation according to the then-known facts. Old journal articles are rarely reexamined in the light of new data. Furthermore, the trend toward highly abbreviated discussions in modern journal articles and ever more complex methodologies and statistical abstractions makes it increasingly difficult to recognize potential errors or to correct them by reexamining raw data. This means that no boxology, even if actually completely correct, will ever be completely compatible with all available data because some of these data are wrong. The theoretical neurobiologist must somehow weight each datum according to the probability that it is correct and then compare theories according to their discrepancies, discounted by the probability of each discrepancy being erroneous. The temptation to simply ignore data because they don’t agree with a pet hypothesis is substantial. Real advancement occurs only as increasing numbers of independent judges arrive at similar conclusions. Unfortunately, popular theories lead to substantial group-think, so the judges are rarely independently minded. It’s a wonder that neurobiology advances at all.
