From “The Neocortex,” edited by W. Singer, T. J. Sejnowski and P. Rakic.
Strüngmann Forum Reports, vol. 27, J. R. Lupp, series editor.
Cambridge, MA: MIT Press. ISBN 978-0-262-04324-3

13

Functional Properties of Circuits, Cellular Populations, and Areas

Kenneth D. Harris, Jennifer M. Groh, James DiCarlo,
Pascal Fries, Matthias Kaschube, Gilles Laurent,
Jason N. MacLean, David A. McCormick,
Gordon Pipa, John H. Reynolds, Andrew B. Schwartz,
Terrence J. Sejnowski, Wolf Singer, and Martin Vinck

Abstract

A central goal of systems neuroscience is to understand how the brain represents and processes information to guide behavior (broadly defined as encompassing perception, cognition, and observable outcomes of those mental states through action). These concepts have been central to research in this field for at least sixty years, and research efforts have taken a variety of approaches. At this Forum, our discussions focused on what is meant by "functional" and "inter-areal," what new concepts have emerged over the last several decades, and how we need to update and refresh these concepts and approaches for the coming decade.

In this chapter, we consider some of the historical conceptual frameworks that have shaped consideration of neural coding and brain function, with an eye toward what aspects have held up well, what aspects need to be revised, and what new concepts may foster future work.

Conceptual frameworks need to be revised periodically lest they become counterproductive and actually blind us to the significance of novel discoveries. Take, for example, hippocampal place cells: their accidental discovery led to the generation of new conceptual frameworks linking phenomena (e.g., memory, spatial navigation, and

Group photos (top left to bottom right) Kenneth Harris, Jennifer Groh, James DiCarlo, Wolf Singer, Jason MacLean, Andrew Schwartz, David McCormick, Pascal Fries, Terry Sejnowski, Kenneth Harris, John Reynolds, Matthias Kaschube, Jennifer Groh, Wolf Singer, Gilles Laurent, James DiCarlo, Gordon Pipa, Pascal Fries, Terry Sejnowski, Gilles Laurent, Martin Vinck

From "The Neocortex," edited by W. Singer, T. J. Sejnowski and P. Rakic.
Strüngmann Forum Reports, vol. 27, J. R. Lupp, series editor.
Cambridge, MA: MIT Press. ISBN 978-0-262-04324-3

sleep) that previously seemed disparate, revealing unimagined mechanistic connections. Progress in scientific understanding requires an iterative loop from experiment to model/theory and back. Without such periodic reassessment, fields of scientific inquiry risk becoming bogged down by the propagation of outdated frameworks, often across multiple generations of researchers. This not only limits the impact of the truly new and unexpected, it hinders the pace of progress.

Outline and Basic Concepts

Two ideas have driven theories of the cortex for decades: the "column" and the "canonical circuit." Although these concepts certainly have a grain of truth to them, it is now clear that they are oversimplifications in need of improvement.

The Cortical Column

The "cortical column" is an anatomical term that connotes a discrete set of cells operating together to perform a computational function. As originally understood, columns are discrete entities, but there can be connections between columns. In this classical view, it must be possible to define boundaries between columns by some means; however, this is almost never possible. Even in the barrel cortex, where "columns" could be easily understood as being defined by the anatomically distinct regions that respond only to single whiskers, boundaries can only really be defined in layer 4; in the other cortical layers, there are no clear cytoarchitectural borders corresponding to the barrels, nor do the cells show sudden transitions in their whisker preference across the cortical surface. In a classical cortical column, features are represented across the surface of the cortex, and in vertical penetrations, the majority of the neurons share the same feature selectivity.

Vernon Mountcastle was one of the first cortical neurophysiologists to emphasize the vertical organization of the cortical layers (Mountcastle 1957). His view of modularity was that a given cortical structure is composed of modules with neighbors of similar functionality. Functionality is determined more by the input a module receives than in the interconnectivity of the cells composing the module. The idea here is that this neighborhood similarity is preserving some type of topology. The basic element of a module is that of a minicolumn, which is about 30 microns in diameter and contains about 100 cells. Larger columns (1 mm diameter) may be composed of hundreds of minicolumns.

Multiple behavioral variables are mapped to the same cortical module, which suggests that these modules can participate in different systems and that this participation is probably dynamic. Mountcastle believed that small clusters of neurons corresponding to columns were fundamental components of cortical function. For each of these cortical processing units, output was produced from structured input. He defined a microcircuit as "the small number

From "The Neocortex," edited by W. Singer, T. J. Sejnowski and P. Rakic.
Strüngmann Forum Reports, vol. 27, J. R. Lupp, series editor.
Cambridge, MA: MIT Press. ISBN 978-0-262-04324-3

of neurons synaptically linked in a processing chain that leads from some particular input…to some particular output" (Mountcastle 1998). These modules were defined almost entirely using criteria such as cellular morphology, layer-specific input-output patterns, and afferent-efferent projection anatomy. He emphasized the importance of discovering the input-output or cortical "operation" of this circuit, considered to be the essence of cortical function, and offered a number of candidates as examples (e.g., differentiation, pattern recognition and generation, coincidence detection, and encoding-decoding of output and input).

To a good approximation, different regions of cortex consist of similar cell types that occur in similar layers, with similar patterns of physiology and gene expression, and are connected using similar rules of connectivity and plasticity. Indeed, recent transcriptomic analysis (Tasic et al. 2018) suggests that while excitatory neurons may be distinct between cortical regions, inhibitory cells are extremely well preserved. Further work to date suggests that the connectivity, physiology, and *in vivo* functions of these cell types are largely preserved between areas (Douglas and Martin 2007; Harris and Shepherd 2015).

It has been over fifty years since the concept of a cortical column was formulated, and we suggest that it needs to be modified to fit current experimental evidence. Rather than a discrete column, the fundamental unit of cortical computation could be described as a "laterally iterated processing unit" (LIPU). Here, the idea is that the synaptic connections of every cell are set up by rules of activity-dependent and chemically hardwired plasticity that are largely independent of the cell's position on the cortical surface. This does not imply that cortical connectivity has to be spatially isotropic. For example, in the visual cortex of carnivorans, connections have a "patchy" appearance that links regions representing similar orientations (Das and Gilbert 1995). Nevertheless, these non-isotropic connections can arise from isotropic rules—a phenomenon familiar in physics (known as symmetry breaking) that allows, for example, a spatially non-isotropic crystal to form even out of spherically symmetrical atomic components.

In the not too distant future, it should be possible to reconstruct connectivity in a cubic millimeter of cortex from electron microscope cross sections (Kornfeld and Denk 2018). This will provide evidence for the patterns of connections between different cell types and the degree to which they are repeated across the cortex. Differences in the spatial scale of repeat distances may be different in different areas of cortex.

As we have defined it, the LIPU is still an example of modularity. What are the boundaries that define this unit? Are they physical (anatomical), computational (e.g., a field of integration encompassed by a "convolutional kernel"), or merely conceptual (a device that makes a complex system easier for scientists to understand)? While the answer is of course still unknown, insights can be gained from computational models of artificial networks, to which we turn to next.

What Does the LIPU Do?

If there is a canonical circuit embedded in the LIPU, presumably it applies a similar processing strategy to diverse types of inputs, performing information processing functions that are useful to the rest of the brain. Of course, we still do not know what this function is but several candidate theories have been put forth. Below, we begin with four of the leading theories.

LIPU Theory 1: Unsupervised Learning

Perhaps the oldest hypothesis for cortical function, the roots of unsupervised learning can be traced to Barlow (1972), Marr (1970), and Konorski (1967). The hypothesis is that cortical excitatory neurons apply unsupervised learning rules to extract features from the data: the input patterns are of distinct statistical structure, which means they will be likely to correspond to features in the natural world of behavioral significance.

At the level of computational models, many unsupervised learning rules have been described. Neural instantiations of standard statistical procedures (including principal component analysis, cluster analysis, and independent component analysis) have all been formalized (Hertz et al. 1991; Dayan and Abbott 2001). The multiple excitatory cell types in the cortex might use different learning rules to instantiate different types of unsupervised learning, and perhaps this is the cause of their characteristically different tuning properties (Harris and Shepherd 2015). Furthermore, cortical cells could represent unsupervised rules that we have yet to imagine.

A simple modification of the theory allows for cortex to implement simple supervised learning rules to form, for example, more detailed representations of stimuli that are present at times of high behavioral salience, as signaled by the activity of neuromodulatory systems (at least in sensory cortex). Substantial evidence suggests not only that cortical plasticity is enhanced by neuromodulators, but that *in vivo* representations of stimuli are stronger and more persistent when neuromodulatory systems are active (Froemke et al. 2007).

This theory is viewed by many neuroscientists as a default. However, it does not account for many experimental facts; for instance, the diversity of inhibitory cells found in cortex and their diverse modulation by nonsensory factors (McGinley et al. 2015b). Indeed, while some unsupervised rules require inhibitory cells for their activity, there are none that need such extreme diversity. Another is the existence of recurrent and feedback excitatory connections, which are not required by such networks.

LIPU Theory 2: Excitatory Recurrence Allows Bayesian Inference

The second theory, which dates at least to computational models in the 1980s (Ackley et al. 1985), is based on the ubiquitous presence of excitatory recurrent connections in cortex. These connections come at a cost: misfunction in recurrent excitatory circuits is at the root of epilepsy. Subcortical structures, while often having recurrent inhibitory connections, do not have such extensive recurrent excitation, nor have recurrent excitatory connections been described in any species other than amniotes. Presumably, therefore, recurrent excitation plays an essential role in cortical computation, and perhaps one that allowed some amniotes to develop impressive cognitive capabilities.

At the heart of this theory is the idea that recurrent excitatory connections encode "priors" or "expectations" concerning relationships between stimuli that the animal is likely to experience. For example, visual scenes often contain extended contours. Recurrent connections between excitatory visual cortical neurons connect neurons whose receptive field centers are elongated parallel to the orientation of the receptive field (Iacaruso et al. 2017). These connections should therefore be able to "fill in the gaps" in continuous contours. Generalizing from this simple sensory example, one might expect recurrent connections to allow associations between higher-order cognitive concepts in multiple cortical regions.

These ideas have been formalized in a computational neural network architecture called the "Boltzmann machine" (Ackley et al. 1985). In a Boltzmann machine, Hebbian plasticity strengthens connections between coactive neurons; if an assembly of neurons is usually driven together by sensory stimuli, connections between them will later enable "filling in" of the activity of neurons whose activity is missing. Mathematically, it was possible to prove that the Boltzmann machine performs a formal process of Bayesian inference: given the available sensory evidence, it estimates possible causes that are compatible with them, sampling an activity pattern from a posterior probability distribution of possible states of the outside world, that are compatible with the available (incomplete) sensory data.

The Boltzmann machine is an attractive model for a function of the LIPU. Furthermore, some behaviors of this network bear an uncanny resemblance to brain activity (e.g., it produces spontaneous activity that mimics the structure of expected sensory inputs). Nevertheless, many features of actual cortical circuits are not required by Boltzmann machines. At least in its initial formulation, there are no inhibitory neurons in a Boltzmann machine, let alone a myriad of cell types. In addition, there is no structured connectivity of different cell types and cortical layers, and there are no spikes. Furthermore, while the original Boltzmann machine was able, in principle, to perform any inference given sufficient time, it learned too slowly in practice to be of use in real-world information processing. More complex versions of the Boltzmann machine architecture, which involve hierarchically repeated populations analogous to a hierarchy of cortical regions, are much more computationally powerful however (Hinton and Salakhutdinov 2006). Thus, a version of the Boltzmann machine that incorporates more complex features could be even more computationally powerful.

LIPU Theory 3: The Liquid State Machine

The third theory, unlike the first two, does not need synaptic plasticity to adapt network connectivity. It is based on the framework of "reservoir" computing (Maass et al. 2002) that uses a randomly structured recurrent neural network to nonlinearly transform a time-varying input signal into a spatial high-dimensional representation. At each time step, the network combines the incoming stimuli with a volley of recurrent signals containing a memory trace of recent inputs. For a network with N neurons, the resulting activation vector at a discrete time t can be regarded as a point in an N-dimensional space. Over time, these points form a pathway through the state space, also referred to as a neural trajectory. This computation serves a feature expansion (i.e., a projection based on many nonlinear basis functions) as well as an explicit implementation of fading memory of past states. While this feature expansion is not specific to a task, task-specific computation is implemented based on learned and weighted task-specific linear or nonlinear mappings of neuronal activity (linear combinations are sufficient).

Even though computational properties of reservoirs can, in principal, have universal computational properties—that is, they can implement any Turing computable function (Buonomano and Maass 2009)—the performance of a reservoir depends on the connectivity, properties of the dynamical elements (i.e., neurons), and the state of the dynamical system—that is, whether the system is behaving regularly, critically, or chaotically (Legenstein and Maass 2007). Moreover, the performance is often much smaller compared to recurrent systems that are optimized (e.g., LSTMs and backpropagation through time) (Hochreiter and Schmidhuber 1997). This gave rise to modified versions of the reservoir computing theory that use neuronal plasticity to self-organize connectivity and the dynamical state of the system (Lazar et al. 2009), implement unsupervised learning to implement noise-robust efficient computations (Toutounji and Pipa 2014), and reward-modulated optimization of reservoirs (Bellec et al. 2019).

There are several features that make this theory of reservoir computing special. The initial reservoir computing idea is useful because it shows that even fully random recurrent networks can compute. This is important from a developmental point of view, since an initially random network can be used to bootstrap the problem and optimize computations over time. This is especially relevant, since the feature expansion into a higher-dimensional representation, carried by a large number of cortical cells, enables random connections to recode information into the "sparse" format helpful to render synaptic plasticity more efficient (Barth and Poulet 2012). Furthermore, because recurrent networks integrate and mix together activity across a range of times, they are able to transform patterns only distinguishable as temporal sequences into spatial patterns, such that a given neuron only fires in response to a specific temporal sequence. This framework has been given a memorable name, describing recurrent random networks as a "liquid state machine."

What the liquid state machine does not do, however, is produce behavior: it just reformats a code into a form that can then be used by downstream structures to learn appropriate behavioral responses. The hard work of learning appropriate responses to stimuli is thus left to downstream structures; the corticostriatal synapse, whose plasticity is well characterized and controlled by dopamine, may be one possible locus for this.

The liquid state machine does not predict the diversity and specific connectivity of different cell types and layers but it is not inconsistent with them. Indeed, computational experiments show that random recurrent circuits with structured connectivity perform better than networks with completely random connectivity (Lazar et al. 2009; see also Singer, this volume). It is therefore conceivable that the complex structure of connectivity found in the cortex evolved to help this function of pattern separation.

LIPU Theory 4: Subtractive Predictive Coding

This hypothesis is in some ways an opposite of the second. A Boltzmann machine amplifies responses to expected stimuli: when an input arrives that matches the types of inputs seen before, the response is stronger, with missing neurons' activity filled in, and more vigorous than it would be to a completely novel type of input. The concept of subtractive predictive coding is the opposite: expected inputs are discarded, while responses to unexpected stimuli are amplified and passed on to downstream structures (Keller and Mrsic-Flogel 2018).

The best example of this processing scheme comes not from cortex, but from the lateral line lobe of weakly electric fish (Bell et al. 1997). These fish sense the surrounding environment by producing electric fields and sensing the disturbances in these fields caused by nearby objects or other organisms. However, most of the electric field impinging on their sensors does not reflect external objects but simply comes directly from their field generation organs, which the fish must subtract out to find the behaviorally relevant external signals. By generating artificial signals as filtered versions of the field which the fish generates, Bell et al. (1997) were able to show that the lateral line lobe performs this subtraction and does so in an adaptive way that also subtracts the signal presented by the experimenters.

The subtractive predictive coding hypothesis posits that the cortex performs a similar function, but it says more: Not only does the cortex subtract simple consequences of one's own actions (such as subtracting the sound of your own voice to hear other people talking over you). It is able to make more complex predictions, for example, computing an expected pattern of visual input based on high-level cognitive expectations, and subtracting it from the actual input pattern to detect subtle features that do not match expectation.

Some very widely observed phenomena can be seen as examples of subtractive predictive coding. For example, presentation of a steady, sustained tone will not cause sustained activity in auditory cortex; it will cause strong activity at its onset and again at its offset ("accommodation"). Given a model where silences and sounds are expected to be sustained, this can be interpreted as producing activity when the times in which a violation of this expectation occurred. Nevertheless, by this standard, accommodation is not a specific function of cortex: it happens in the sensory receptors themselves and again at many levels of the processing hierarchy. It may be that the cortex specializes in subtracting predictions of advanced statistical models of the outside world, but the experimental evidence for this is mixed. For example, Keller et al. (2012) reported that ~10% of neurons in mouse visual cortex respond to mismatches between self-motion and visual motion signals, whereas Saleem et al. (2013) reported that visual cortical neurons responded instead to a match between these two signals, which would be more consistent with hypothesis 2 than hypothesis 4.

Spontaneous Activity

Another feature of cortical physiology which we refer to is *spontaneous cortical activity*. Clearly, it should come as no surprise to find spontaneous activity in the nervous system. If there was no spontaneous activity in the circuits controlling respiration, we would have a problem. However, the presence of spontaneous activity in the sensory systems is more surprising. Spontaneous activity in sensory systems, and the related phenomenon of variable responses to sensory stimuli, seem fairly specific to cortex: much lower levels of variability are seen in subcortical mammalian structures. As yet, there is no consensus on the function of structured spontaneous cortical activity, but it is possible to list some hypotheses, again non-exclusive.

Spontaneous Activity Theory 1: Nothing, or Worse

The first possibility, which cannot be excluded based on current data, is that spontaneous cortical activity serves no function at all. The cortex is spontaneously active under anesthesia, and as far as we know performs no information processing in this state. Although spontaneous cortical activity costs some energy, it may be that this is so minor, in evolutionary terms, that an animal suffers little disadvantage, even if there is no need for it to occur at all.

An even more extreme view holds that spontaneous activity is worse than useless: it is a form of noise that actually impairs processing of sensory inputs by interfering with neuronal representations. In this view, neurons are noisy devices, and worse, this noise becomes correlated through the cortex's highly recurrent connectivity. One result that could be taken as evidence for this perspective is that correlated fluctuations in primate visual cortex get smaller when the subject is attending to a sensory stimulus (Harris and Thiele 2011).

An alternative interpretation of this result is discussed below. A related concept from motor neurophysiology is that neurons in the motor cortex can merge together such that their combined activity is a "null space" that is occupied specifically when muscle activity is absent (Kaufman et al. 2014).

Spontaneous Activity Theory 2: Imagery, Memory Recall, and Consolidation

Spontaneous activity shares many features with sensory-evoked activity. For example, Kenet et al. (2003) have reported similarities between sensory-evoked and spontaneous activity patterns in anesthetized cats and, recently, consistent observations were made in awake ferrets, with spontaneous activity being more exuberant in the awake than in the (lightly) anesthetized state (Smith et al. 2018). Thus, one might hypothesize that spontaneous activity in sensory systems correlates with processes such as imagery and memory recall.

In this view, the brain spontaneously produces patterns of neural activity that mimic actual sensory responses and have similar consequences on downstream structures. These consequences might involve production of actions: for instance, when remembering the nature and location of an object currently hidden from view, the brain might reproduce activity patterns similar to those the object would itself produce, thus allowing motor actions to be performed similar to those the object would itself produce.

Even if a spontaneous activity event does not directly produce action, it can have other consequences, such as changes in synaptic strengths. For example, recapitulation of activity patterns that occurred in previous behavior could cause further consolidation of the synaptic changes that encoded this memory; consistent with this view, interruption of spontaneous events in hippocampus after behavioral experience disrupts formation of long-term memories of that experience (Girardeau et al. 2009; Jadhav et al. 2012). More complex possibilities exist: spontaneous activation of neuronal assemblies containing overlapping cell populations could cause changes in synaptic strengths linking these neurons, thereby forming associations between previously unrelated concepts. This process could be a basis for the process that humans subjectively describe as "thinking."

Spontaneous Activity Theory 3: Nonsensory Context

While spontaneous activity shares some structural properties with sensory responses, they are far from identical (Scholvinck et al. 2015). Perhaps, then, a major function of spontaneous cortical activity in sensory systems has no direct connection to sensory processing, but instead encodes nonsensory variables, which are integrated with, and can modulate the detection of, sensory stimuli (McGinley et al. 2015a).

An important clue to this comes from the mouse visual cortex. Activity in visual cortex changes when mice run, even in complete darkness (Niell

From "The Neocortex," edited by W. Singer, T. J. Sejnowski and P. Rakic.
Strüngmann Forum Reports, vol. 27, J. R. Lupp, series editor.
Cambridge, MA: MIT Press. ISBN 978-0-262-04324-3

and Stryker 2010). This activity presumably has nothing to do with expected sensory stimuli. Furthermore, neurons in sensory cortex respond to rewards (Shuler and Bear 2006), and imaging of axons arriving in sensory cortex from elsewhere shows they convey very complex nonsensory information. It may be that spontaneous cortical activity is in fact a high-dimensional representation of an animal's current cognitive and behavioral state, which the cortex integrates with sensory information (Stringer et al. 2018). The optimal behavior to produce in any circumstance depends on a combination of sensory input and internal context; by integrating these two classes of information, the cortex may provide information allowing an animal to perform behaviors that integrate sensory and nonsensory data.

Cortical traveling waves, which have been observed in both sleep states and awake state, are another source of spontaneous activity (Muller et al. 2018). They modulate the membrane potentials of neurons in a spatially organized way and vary in frequency from theta (4–8 Hz; Lubenov and Siapas 2009) to gamma (30–80 Hz; Gabriel and Eckhorn 2003).

Spontaneous Activity Theory 4: Housekeeping / Homeostasis

Our final theory suggests that spontaneous activity is not a reflection of information processing per se, but rather that it functions to maintain the biophysical and biochemical state of the network. Spontaneous electrical activity is prominent in the development of the nervous system from the earliest stages (Spitzer 2006). The function of this early spontaneous activity presumably has nothing to do with processing of sensory information, memory recall, or motor variables. Instead, it seems to function to specify neural circuits, for example, to determine the differentiation, migration, and wiring of developing neurons. Cortical spontaneous activity shows very sudden changes with development: adult patterns show a substantially different structure to those present earlier in development (Luhmann and Khazipov 2018). Nevertheless, it remains possible that cortical spontaneous activity in adults plays at least a partial role, similar to its role in early development, enabling and guiding low-level maintenance of cellular and circuit properties. Several studies have reported that axonal conduction delays in the cerebellum are tuned to allow complex spikes from the inferior olive to arrive in a precisely timed manner, despite differing physical lengths of these axons (Sugihara et al. 1993; Baker and Edgley 2006). If this is the case, some homeostatic mechanism must enforce these constant delays; spontaneous activity, perhaps during sleep, could be a key part of the process. Spontaneous activity during sleep has also been proposed to enable "downscaling" of firing rates and synaptic strengths built up during waking (Tononi and Cirelli 2014) or other metabolic functions (Vyazovskiy and Harris 2013). Enabling these low-level metabolic and circuit functions might be a key function of spontaneous activity in both waking and sleep, parallel to the information-processing roles described above.

Considerations from Evolution

The brain of any species cannot be understood in isolation but is best considered in an evolutionary context. Two critical concepts related to evolution, in general, and the brain, in particular, are *inheritance* (i.e., features that have been continuously present in a given phylogenetic lineage) and *convergence* (i.e., features that arose independently multiple times but which accomplish similar functions).

For an example of inheritance, consider the molecular building blocks of nervous systems (e.g., ion channels), which can be found in a highly similar form in bacteria. The synaptic transmission machinery operates with the same molecular components and principles across all animals, as far as we know. In fact, many of the molecular elements and their functional interactions were worked out in yeast. Sponges have some cells, called flask cells, that contain many of the molecular components of the postsynaptic compartment (ionotropic receptors, e.g., are missing but metabotropic ones are present). Flask cells, however, are not neurons, and sponges have no nervous system or synapses (Sakarya et al. 2007). So, either synapses evolved by borrowing and adapting already existing components, or present-day sponges lost a nervous system that existed in one of their ancestors. Sponges diverged from us and other animals some 600 million years ago.

Short-term synaptic plasticity mechanisms, such as facilitation and depression, are found in simplest nervous systems. Spike timing-dependent plasticity exists in insect nervous systems but whether these use glutamate and NMDA receptors is not currently known. Synchronization has been discovered in mollusks, insects, etc. Spatiotemporal representations are found in invertebrate sensory systems (e.g., in locust olfaction; Wehr and Laurent 1996; Mazor and Laurent 2005), or leech motor and premotor systems (Briggman et al. 2005).

Examples of convergence include looming sensitivity in single neurons in insects and birds. It consists algorithmically as a division of angular velocity by an exponential of angular size (Gabbiani et al. 2002). This algorithmic description also applies to cells in thalamic nucleus rotundus in diving birds (Sun and Frost 1998). It is very unlikely that the same computation (or need) existed in their common ancestor, which was some sort of worm. Another example is Jayaraman's result (Seelig and Jayaraman 2015) on head direction-like cells in the central complex of insects, which is similar to models of head direction cells in mammalian hippocampus. The olfactory system is also interesting: despite the fact that the molecular nature of the olfactory receptor genes is different in invertebrates and mammals, the organization (convergence to glomeruli, divergence and random-like distributed projections to second structures—piriform cortex or mushroom bodies) is similar, probably through convergence.

Evaluating homology across species from very different lineages is critical for cross species comparison but is a challenging task. For example, a dorsal telencephalon or pallium is part of the vertebrate brain *bauplan*. Thus it can

From "The Neocortex," edited by W. Singer, T. J. Sejnowski and P. Rakic.
Strüngmann Forum Reports, vol. 27, J. R. Lupp, series editor.
Cambridge, MA: MIT Press. ISBN 978-0-262-04324-3

be found in fish, amphibians, reptiles, birds and mammals. To trace back the evolution of the mammalian cortex, one has to look first at the outgroup of mammals; that is, reptiles.

Unlike fish and amphibians, a large portion of reptilian pallium has a three-layered organization, indicating that a layered cerebral cortex emerged about 320 million years ago in the ancestor of mammals and reptiles (the amniote ancestor). In addition, reptiles and birds harbor a nonlaminated pallial region, called dorsal ventricular ridge (DVR), where neocortical-like circuits have been identified.

The structural and functional differences of reptilian and mammalian pallial regions have fueled controversies on the evolutionary origin of the mammalian neocortex. How can we compare reptilian and mammalian pallial regions, cell types, and circuits? Do similarities result from homology or convergent evolution? And how can this discussion inform us on the evolution of cortical function?

Homology hypotheses can be tested by comparing early development, gene expression, and connectivity. The existence of thalamo-recipient neurons in the anterior DVR led to the "equivalent circuits" hypothesis, stating the homology of anterior DVR and neocortical L4 neurons. The analysis of a small set of molecular markers supported this idea. However, anterior DVR and neocortex develop from two distinct regions of the embryonic pallium, and the conservation of developmental fields would predict the homology of anterior DVR with mammalian claustrum and parts of the pallial amygdala, derived from the ventrolateral pallium.

To test these hypotheses further in an unbiased manner, Molnár and colleagues compared gene networks in micro-dissected chick and mouse pallial regions (Belgard et al. 2013). Their results show that only five genes are shared between a L4 gene module and an anterior DVR module. Micro-dissected brain regions, however, may contain cells of different types in different proportions, and this might confound the analysis and hide similarities. To overcome this limitation, Tosches et al. (2018) applied single-cell RNA sequencing to the turtle and lizard pallium.

The single-cell approach allows the analysis of small and sparse cell populations such as cortical interneurons. The comparison of turtle and mouse data shows that the same classes of GABAergic interneurons exist in the two species: interneurons derived from medial (MGE) and caudal ganglionic eminences (CGE), including somatostatin, parvalbumin-like and vasoactive intestinal polypeptide-like types. This suggests that developmental and/or functional constraints led to the conservation of these interneuron types for over 320 million years.

High-level clusters of glutamatergic neurons map to distinct regions of the reptilian pallium: the hippocampus, dorsal cortex, olfactory cortex, the so-called "pallial thickening," and the DVR. These regions express different combinations of transcription factors, reflecting their distinct developmental and evolutionary histories. The comparison of regional transcription factor codes in reptiles and mammals supports the hypothesis that the anterior DVR is homologous to the mammalian lateral amygdala, as also indicated by the fact that these regions develop from homologous developmental fields and establish similar connections with the rest of the brain. Nevertheless, many effector genes (e.g., ion channels, cell adhesion molecules) are shared between the reptilian anterior DVR and the mammalian neocortex, indicating that the expression of the same gene sets in these two pallial regions is regulated by different transcription factors. In conclusion, different pallial regions expanded independently in the reptilian and mammalian lineages—ventral pallium (anterior DVR) versus dorsal pallium (neocortex)—resulting in the convergent evolution of gene expression and circuits.

Representations and Neural Codes

What Is a Code, Anyway?

Encoding and Decoding Models: Definitions and Scientific Goals

The terms *code*, *representation*, *encode*, and *decode* have become highly overloaded in neuroscience: different people use the same phrase to mean very different things, so that investigators often talk past each other rather than coming together to synthesize and integrate ideas. Grounding of these terms requires discussion of the goals and the assumptions in the models used to achieve those goals. A subset of our group had a lively discussion on these points, and here we attempt to explain those sometimes divergent viewpoints.

It is widely assumed that neural spikes are, for most problems of interest, the carriers of information to support moment-to-moment behavior. (Here "behavior" is broadly construed to include sensation, cognition, and action, and could be studied in an ethological or a laboratory context.) Under that assumption, three main types of data are typically measured and/or experimentally controlled:

1. energy patterns that impinge on sensory epithelia,
2. spike patterns in populations of neurons in one or more locations in the brain, and
3. the positions of one or more parts of the body (e.g., arms, eyes, vocal apparatus).

The goal of much of systems neuroscience is to use such data to "understand" how the internal parts of the system operate together to execute complex sensorimotor loops (i.e., "cognition," "complex behavior," "intelligence," etc.). A more modest goal may be to describe the information content contained in a population of neurons, without assumptions of the explicit role these neurons may play in the behavior generation. This more relaxed approach may mitigate many of the arguments between differing viewpoints that arise from invalid assumptions. Nonetheless, experiments define concepts derived from such measurements (e.g., "motivation," "reward expectation"), and it is important to keep in mind that such definitions are not direct measurements; they are only inferences, as they assume one or more underlying models of what the brain is doing. Indeed, all such assumed internal latent variables are inferred from the same three basic measurements above: stimuli, neural activity, and behavioral measurements. And, if judged at all, each model is judged on the accuracy of predictions it makes for other observed variables (typically neural spikes and/or behavior).

The form of the understanding we seek is not usually explicitly stated. We argue, however, that it should ideally be in the form of inferred, neural mechanistic causal models that describe the linkages between those three types of measurements. A *neural mechanistic model* is a model that minimally contains approximations of neurons and their connections. A *causal neural mechanistic model* is one in which external perturbations can be injected or model parts removed, so that the resulting effects on the other parts of the model will be accurately predicted.

As a point of departure, we may begin to understand some aspects of the transformations taking place as raw sensory signals propagate through the nervous system. These transformations are usually considered mechanistically; that is, how an output of some entity (e.g., a neuron deep in the visual system) fires relative to an input (e.g., a pattern of light energy on the retina). As an example of a neural mechanistic causal model, consider a transfer function that is implemented as a set of modeled neural elements and their connections, which aim to describe and predict this transformation accurately. This model can (a) explain how the stimulus is responsible for the output, (b) make predictions of what other internal neural responses should be found along the way, and (c) predict how direct perturbations of those internal elements will lead to perturbations in the output. While such predictions may turn out to be incorrect, the model can drive a principled selection of future experiments, which would aim to reject this model in favor of one or more alternatives. We refer to this model as an "encoding model" (see Figure 13.1). An encoding model has the advantage of providing a concise description of the relationship between neural activity and the variables being encoded, but it may turn out that cause-and-effect relationships might be very complex in biological systems. Still, until someone proposes another way to make scientific progress, important work continues utilizing this framework in the hope that such complexity can be overcome.

This conceptual discussion and the three types of measurements listed above (stimuli, neuronal spikes, and behavior) lead one to see that there are two primary types of neural mechanistic causal models:

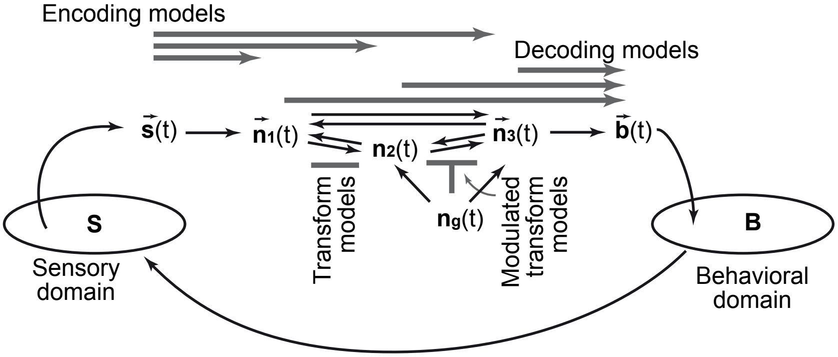

Figure 13.1 Nomenclature for different types of modeling paradigms of mechanisms underlying moment-to-moment sensory-cognitive-motor behavior. If we limit ourselves to spiking activity and moment-to-moment behavior, four modeling paradigms may encompass potentially all of the model-building activities in the field. **S** represents a defined sensory domain (e.g., the set of all possible visual movies of a specified size and resolution); **s**(t) is a sample from such a set (e.g., a frame of one movie). Each **n**(t) is a potentially time-varying population vector of neural activity in a brain region (e.g., the entire set of pyramidal neurons in layer 2/3 of primate visual area V1 that project to primate visual area V2), whereas **b**(t) is a potentially time-varying vector that describes, e.g., the current positions of all body parts, which itself lives in a set of possible configurations of body parts (**B**). Of course, reduced description **b**(t) is also possible and potentially useful (e.g., the subject's choice on each behavioral trial). Arrows show the neural populations that are assumed to have direct connections: behavior can influence the environment and thus the next impinging sensory sample (outer loop). The two most dominant forms of model building are (a) encoding models, which are causal neural mechanistic models that apply to **S** to generate predictions of the responses of neuron populations (**n**) and (b) decoding models, which are causal neural mechanistic models that apply to one or more specified **n**'s to generate predictions in a specified behavioral domain (**B**). Although the examples depicted above are for visual-cognitive-motor domains, similar examples can be readily defined for other sensory-cognitive-motor domains.

1. Causal neural mechanistic models that link between the sensory epithelia and neural spiking activity at one or more places in the nervous system (which we refer to "encoding models").
2. Causal neural mechanistic models that link between neural spiking activity at one or more places in the nervous system and the positions of the body parts (here termed "decoding models").

While much progress has been made on encoding models, disagreement surrounds decoding models: what they are, their utility, and how they should be interpreted. Here we emphasize that the goal of the decoding approach is to gather knowledge of the natural causal brain processes, not simply to control an externally attached device. We also emphasize that science must propose hypotheses that can be implemented (through models) so that they can be tested; that is, a downstream homunculus is *not* a scientific decoding model.

Broadly speaking, "neural decoding" is any implementable analysis that demonstrates the ability to predict some outcome by extracting information from neural recordings. The implementation is usually in the form of some kind of extraction algorithm (e.g., linear filter, population vector algorithm, Kalman filter, or deep learning neural network) and the success of the decoding analysis is measured by its accuracy of predicting the targeted outcome (e.g., arm trajectory, behavioral choice). *An above chance decoding performance shows that information about the predicted variable(s) is present in the recorded neural activity.* This has scientific value because information is exposed that may not be obviously present in individual neurons and because information identification constrains the hypothesis space of causal neural mechanistic models that may exist between the recorded neurons and the predicted behavior (i.e., "decoding models") (Majaj et al. 2015).

A common point of confusion in our field is how the modeler views the decoding analysis: whether the extraction algorithm itself is physically realized, in this case, within the nervous system (perhaps within a particular anatomical location or set of locations), or whether it is merely a tool used to identify the information content at a node defined by the neurons whose activity patterns were recorded as input to the extraction algorithm. Again, as defined above, decoding means extracting information from neural firing rates. This can be performed within the nervous system or by an outside observer. In either case, the code itself exists, independent of whether or not it is decoded.

Based on the principle of decoding, we hope to better understand the information contained in patterns of neural activity. Decoding analyses expose information that is transformed by the brain as it propagates through the system. This allows scientists to propose and construct models and to make predictions about the behavior (which the decoding test defined above has already at least partly achieved) as well as about neural activity that intervenes between the originally recorded neurons and the behavior. Similarly, scientists working on encoding models should propose alternative causal linkages from sensory epithelia to patterns of neural activity and conduct experiments to distinguish between those alternative encoding models. Of course, all models are wrong at some level, so that scientific work will continue indefinitely, at least until causal neural mechanistic models of sufficient accuracy and predictive power are obtained to support the application goals of the society that funds the scientific research (e.g., new ways to intervene causally in the system to ameliorate brain disorders). In what follows, we formalize these definitions a bit further in the hopes of grounding these key ideas.

We define a potentially time-varying, multidimensional stimulus in the external world as $\mathbf{s}(t)$, where any specific $\mathbf{s}(t)$ is an element from a domain $\mathbf{S}$ (e.g., a set of natural images, movies, sound sequences). We define a potentially multivariate, potentially time-varying body state in (lagged) response to an element drawn from the domain $\mathbf{S}$ as a behavior $\mathbf{b}(t)$ that will reside in the range $\mathbf{B}$ (e.g., the possible positions of the body parts that are chosen to be monitored).

In addition, we define the time-varying activity vector of a given subset of neurons (e.g., all layer 2/3 pyramidal cells in visual area V1 that project to layer 4 of V2) as **n**(t). For instance, we can imagine the following example population vectors: **n**[V1](t), **n**[IT](t), **n**[basal ganglia](t), etc. Given this framework, we propose the following possible model paradigms, which must be built and tested (Figure 13.1):

- encoding models, which map from **s**(t) to some **n**(t),
- decoding models, which map from some **n**(t) to **b**(t),
- neural population transform models, which map between **n1**(t) and **n2**(t), potentially recurrent, and
- modulated neural population transform models, which map between **n1**(t) and **n2**(t), subject to another neural population(s) **ng**(t).

We do not aim to be overly prescriptive, but to engender shared scientific terminology and associated goals.

Progress has resulted from, and some of us believe that it will continue to result from, making measurements and using those to estimate the parameters of different hypothesized model families in each one of these modeling paradigms. We note that this framework is highly general and inclusive. For example, it includes the notions of external time and the notion of recurrent processing (note the arrows in Figure 13.1). Our goal here is not to strongly limit the alternative model families that might be considered, but to define the expected inputs and the expected predictors of any actual model. Once a model paradigm is chosen (from the above list), and the parameters of a model are determined (i.e., the model is selected from a hypothesized model family), the accuracy of the model is evaluated by its ability to predict its output variable(s) from its input variable(s) on held-out input variable settings (i.e., values of the input variables that were not used to fix the model parameters). That is, the test is generalization within the domain (e.g., **S**) of the model, where generalization can be defined as tests of increasing distance from the inputs used to determine the parameters of the model (aka training data). In addition to its predictive accuracy (generalization), the model might also be judged by its elegance or simplicity (e.g., amount of training data needed to specify the model parameters), the degree to which the model components correspond to known neuroanatomy, and/or the minimum description length of the model.

For a given stimulus domain (**S**) and behavioral domain (**B**) and proposed intermediate (hidden) neural population vectors, nearly all models in systems neuroscience can be placed into one of these paradigms: *encoding*, *transform*, *modulated transform*, *decoding*. Note that this is a very general and inclusive framework for common ground, as it leaves open the questions of timescales, dimensionality, model families to be considered, etc. The definitions are also broad enough to include instantiated neural network models of all existing potential normative theories: predictive coding, generative models, Bayesian inference, etc. Each of these choices might be highly specific to an area of conceptual study, though it is hoped that groups working in similar domains (e.g., recurrent sensory systems) might build and test some common model families so that model families that naturally work well (highly accurate in predictions) in more than one sensory system (albeit with potentially different model parameters in each case) will be readily discovered.

To fit such a model, parameters do not need to be estimated entirely via empirical data (e.g., neural responses and/or behavior). Indeed, the currently most accurate (predictively accurate) models of the primate ventral visual stream were discovered by fitting the parameters of a global model that includes each of these model types: encoding, transform1, transform 2, transform 3,… transform 6, decoding. Notably, the parameters of each and every one of those component models were fit entirely from a large number of pairs **(s**, **b)** drawn from a large domain **S** (images in the central ten degrees) and **B** (category reports). The internal "neural" populations of these deep artificial neural network models turned out to be highly predictive (over new **s**'s taken from **S**) of the responses of the internal neural population vectors (**n**) at multiple stages of the ventral stream (**n**[IT], **n**[V4]) (Yamins et al. 2013, 2014) and **n**[v1] (Tolias et al. 2001). Critically, that success was enabled because the mapping from **S** to **B** was chosen in a way to make the tasks computationally challenging (invariant object categorization). This thus appears to be an example of convergent evolution—artificial neural networks for visual processing have "evolved" (under human organized optimization pressure) to have internal representations that look very much like the brain's internal representations. Stepping back, the organizing point is that models which map from **S** to **n** are all examples of *encoding models* (above), and the goodness of these models is not judged on just capturing data, but on predicting new data. Such models are still far from complete, even in the visual system.

We note that encoding models that take high-dimensional sensory data as input naturally contain notions of receptive fields (RF, i.e., fields can be measured by doing virtual electrophysiology on the model, or by tracing the connections in the model). However, these encoding models are much more accurate (at predicting their population vectors (n) than the basic RF encoding model (also see below on the limitations of tuning curves). Thus, while the concept of a RF is still a useful teaching concept that can predict some aspects of any neuron's response, it is not a very accurate encoding model.

In that regard, we note that current encoding models are still not able to explain all aspects of the neural responses, most notably, many current encoding models lack temporal dynamics. And modulatory transform models have not yet been incorporated in any serious way. Testing new encoding models in that expanded family of recurrent artificial neural networks is ongoing work in multiple groups (i.e., causal neural mechanistic models), and we hope that those model families will be informed by the discussions and the data presented at this Forum.

It is useful to consider encoding and decoding in terms of statistics as well as from a communication/information perspective. As an example, consider the classic behavioral paradigm of locomotion. Suppose we want to predict the phase of locomotion based on the firing rate of a single, isolated afferent fiber from a cutaneous pressure receptor in the skin of the foot (Werner and Mountcastle 1965). If we assume a quasi-steady state (e.g., walking at a constant speed on a treadmill), we should be able to make a step-phase prediction with some degree of accuracy. If the afferent fired a graded volley of action potentials that appeared as a single bump in an event-triggered histogram beginning 10 ms after foot-strike, one could try to match an instantaneous firing rate to a phase of the step cycle. However, even if the firing rate pattern was exactly the same for each step, that prediction would be uncertain, because the same firing rate occurs twice for each step (except at the top of the bump). The histogram can be thought of as a probability distribution (probability of firing at a time point during the step cycle). So, acting as an ideal observer under ideal conditions, you would have a 50% error rate. In more realistic situations, where the behavioral state is not limited to isolated walking on a treadmill, that skin stretch receptor is being activated continuously in many different behavioral contexts. Predictions of behavior based only on observing that afferent's firing rate would be very poor.

This thought experiment was based on a single primary afferent. In general, neurons are driven by many different sources. For example, neurons in the motor cortex have been shown to carry information about ten different components of arm and hand motion simultaneously (Wodlinger et al. 2015). These components include the velocity of the hand as it moves through 3D space, orientation of the wrist, and shaping of the fingers. Because of the high-dimensionality of this encoding, the same discharge rate is associated with many different weighted combinations of these parameters. Fortunately, better decoding (i.e., better ability to predict action) can be achieved by observing the firing rates of many neurons simultaneously. Even though parameter encoding by individual neurons is redundant, the bias of each neuron's firing rate to a particular combination of these parameters (i.e., its "tuning") is specific. This uniqueness of a neuron's tuning function makes it possible for extraction algorithms to decode parameters encoded simultaneously in single-unit activity. Furthermore, parameters that are weakly encoded by individual neurons, but have a consistent effect on the firing rate of many members of the population, can be extracted with these algorithms. This general principle is why populations of neurons are needed for more successful prediction of behavior from neural population.

While successful prediction of behavior (i.e., "decoding" by an observer) can add support to the inference that the recoded neurons are causally linked to the behavior in a neural mechanistic manner, such success does not guarantee that it is the correct inference. How might that inference be strengthened? We see two ways: First, requiring an ever more detailed prediction of the specified

From "The Neocortex," edited by W. Singer, T. J. Sejnowski and P. Rakic.
Strüngmann Forum Reports, vol. 27, J. R. Lupp, series editor.
Cambridge, MA: MIT Press. ISBN 978-0-262-04324-3

behavioral domain (B, Figure 13.1) should tend to show the limitations of incorrect causal inferences (i.e., incorrect decoding models as shown in Figure 13.1). This cannot prove the causal link, but it could still lead us to the correct neural mechanistic model that intervenes between the neural activity and the behavior. Second, direct perturbation of the recorded neural elements (e.g., silencing of multiple, targeted individual neurons) should produce behavioral effects that are precisely predicted by the decoding model. Ever more detailed perturbations can again be used to lead us to the correct neural mechanistic decoding model. Both approaches should ideally be applied to the brain subsystem under study. The larger point to keep in mind is that decoded information emerging from the processing of neural data, no matter how accurate it may be, does not guarantee that this information is used by the nervous system. It is only a starting point to a very large family of neural mechanistic causal models, and those causal relationships—and thus the most veridical models—are likely to be highly complex.

We can illustrate this type of problem with the walking experiment. Within this task are a series of behavioral events (e.g., a trigger cue, onset of movement, target acquisition, reward administration). A peri-event histogram of the skin afferent's firing rate, triggered on one of the behavioral events, will have a structure (e.g., a bump) representing the probability of the afferent having an action potential at a point in time relative to the event. In terms of communication theory, because there is a correlation between the event and the firing rate, there is information (reduction in uncertainty) being transmitted between the event and the neuron's firing rate. This relation can still be very noisy and may not mean that there is any kind of direct synaptic connectivity (direct causal relation) between the event and the change in firing rate. This is an important point, has led to a great deal of confusion, and is often contentious. Walking is a cyclic behavior in which the entire body oscillates with a period equal to the step-cycle length. Every part of the skeleton, every somatic pressure sensor, visual and auditory input, almost all muscles, and probably most neurons are going to be entrained by this periodic behavior. A histogram of almost any neuron's firing rate will show some kind of structure. That neuron is therefore transmitting information encoded as firing rate through the step cycle even though it is unlikely to be linked by any direct "circuitry" to the neural source driving locomotion. The foot receptor transduces pressure and, when activated by stretched skin, is "causing" the neuron to fire, but whether the signal in our afferent fiber is transmitted in a way that is decoded subsequently is unknown. Simplistically, we can record the firing rate from neurons in many other parts of the nervous system and those rates will be highly correlated to that of our afferent. This in no way means that our afferent is "causing" those other neurons to fire.

Given the definitions of encoding models and decoding models above, then a neural "code" is conceptually defined as a particular measure of information somewhere in the brain that is *both* a product of an encoding model and an

From "The Neocortex," edited by W. Singer, T. J. Sejnowski and P. Rakic.
Strüngmann Forum Reports, vol. 27, J. R. Lupp, series editor.
Cambridge, MA: MIT Press. ISBN 978-0-262-04324-3

input to a decoding model. In this formulation, the job of alternative decoding models is to specify *what* measure of information is causally critical (to a domain of behavior) and *how* it is causally critical (to that domain of behavior). And a key job of alternative encoding models is to explain and predict *that* information using neural mechanistic causal linkages from sensory epithelia to the specified measure of neural population activity. That is, a putative neural "code" is specified with respect to the knowledge of (or at least the hypothesis of) the downstream causal circuitry. In the next section, we provide another perspective on neural codes from the perspective of communication theory, which also relies inherently on the notion of a combined encoding and decoding model.

Coding from a Communication-Theoretic Perspective

The concept of a "code" is one of the most commonly used terms in the neurosciences. Outside the neurosciences, we usually mean something very specific when we refer to a "code" or "coding"; these concepts have been formalized in communication and information theory. In the neurosciences, we have often a much more rudimentary notion of a "code": Does the communication theoretic notion of coding make any sense for the brain, and to what extent? What are the gaps in our knowledge about the "neural code"? Looking at various levels in the nervous system (retina, V1, M1) we see that in some cases the communication theoretical approach is sensible (retina), in others it is by approximation (V1), and in others becomes highly problematic (M1). Is it time to abandon the notion of a "code" because it erroneously carries with it all the communication theoretic baggage and the notion of "representation," or can we use this baggage?

Using concepts of information and communication theory, we can specify a number of conditions for coding in the classical communication theoretical sense and differentiate this from a more rudimentary concept of a code in the sense of merely containing information. We will distinguish two types of codes:

- Code Type 1: A code in the communication theoretical sense, which we call a *coding algorithm*.
- Code Type 2: A code that depends on the human observer, treating the link between the encoded variable and neural response as a virtual communication channel. A code in this sense means to contain information or, in other terminology, to have a tuning curve (see below).

There are five conditions for a Type 1 code in the communication theoretical sense:

- Condition 1: The sender receives some data.

- Condition 2: The sender encodes this data with a series of symbols (e.g., bit sequences), a "code," in a systematic manner (according to an algorithm).
- Condition 3: This achieves some transmission or storage goal.
- Condition 4: This code is constructed in a way that achieves data compression and allows error correcting at the receiver site (to solve the two practical problems that bandwidth is limited and communication channels are noisy).
- Condition 5: Finally, there needs to be some receiver (decoder) that can "understand" and do something meaningful with the transmitted symbols (e.g., decode the original signal or perform some action based on the message).

If such a model applies, then this permits us to use a powerful toolbox of techniques from communication theory to analyze the system. It also allows us to think about representations as a coding algorithm as well as to better understand other computational frameworks, like deep neural networks (Shwartz-Ziv and Tishby 2017). The definition of a Type 1 code makes clear that for the nervous system, we need both the notion of a "receptive field" and a "projective field" (Lehky and Sejnowski 1988); in our terminology used earlier, we need both an encoding and a decoding model. Below, we will further examine how the concepts of a Type 1 code apply to the nervous system (visual and motor system) and evaluate what limitations and gaps in our knowledge exist.

Usually, when neuroscientists talk about a "code" they take a stimulus (S), some neural activity (R), and compute a mutual information function I(S;R) (or any other measure of dependence) to demonstrate that R "encodes" S. However, this is not equivalent to a demonstration of the existence of coding in the communication theoretical sense; it merely provides a generic measure of statistical dependence and shows that R contains information about S (Code Type 2). The fact that neural activity contains information does not mean that this information is being used or that the information can be easily decoded. Nevertheless, the exercise of quantifying whether neuronal populations contain information and which algorithms work best to decode generates very useful hypotheses about whether these neuronal populations may "encode" in the Code Type 1 sense and which algorithms the neural decoder may use. This is also important for constructing brain–computer interfaces (Schwartz et al. 2006).

Coding in the Visual System

It has proven fruitful to model responses of photoreceptors or neurons in the retina as encoding the image that falls on the retina: The retina transforms the received image (Condition 1) in another signal, in a systematic way (electrical impulses/currents, organized topographically) (Condition 2), to achieve some goal (transmission of information about this image to the cortex over a channel with limited bandwidth and some noise) (Condition 3). The code is constructed according to some smart principles, achieving data compression (e.g., removing redundancies between pixels; Schwartz and Simoncelli 2001) (Condition 4), and there are receivers that understand the message and do something meaningful with it; namely the superior colliculus and the LGN, ultimately leading to behavior (Condition 5).

Whether and/or how the concept of coding applies to cortical areas like primary visual cortex or area IT remains far from clear. First, what does V1 encode (Condition 1)? Area V1 does not only encode the image on the retina but is sensitive to many other internal and external variables, such as arousal, movement, attention, and other sensory modalities (McAdams and Maunsell 1999; Niell and Stryker 2010; McGinley et al. 2015b; Stringer et al. 2018). In addition, V1 might not just encode the image on the retina but perform a type of Bayesian inference about the causes of the sensory data using priors and expectations (Rao and Ballard 1999; Friston and Kiebel 2009).

Second, how does V1 encode the data (Condition 2)? The population rate vectors are commonly assumed to form the coding substrate. However, it has not been demonstrated that they yield a "complete" representation of the image on relevant behavioral timescales (see, e.g., Van Rullen and Thorpe 2001; Resulaj et al. 2018), and there is evidence that there is additional stimulus information encoded in spike timing (discussed further below). Furthermore, we do not know which spikes are part of the code that is transmitted to other areas and which spikes are merely part of the coding process (e.g., spikes from interneurons), nor do we precisely understand the role that correlations play in coding. To make matters more complicated, there is an abundance of spontaneous ("dark") activity (discussed further below) that does not seem to encode any sensory information, and there exists tremendous state variability in sensory responses (Harris and Thiele 2011; McGinley et al. 2015a).

Third, what are the coding design principles in area V1 (Condition 4)? There is evidence that V1 receptive fields are optimized for sparse coding, and processes such as surround modulation have been interpreted from the perspective of efficient coding (Olshausen and Field 1996; Rao and Ballard 1999).

Fourth, who is the receiver of V1 information and what does this receiver do (Conditions 3 and 5)? There are many receivers of V1 information, including cortical areas (V2, V4, MT) and subcortical areas (e.g., cerebellum, striatum, superior colliculus). Cells may transmit different information depending on cortical/subcortical projection targets (Lur et al. 2016). Furthermore, it remains largely unclear which information in V1 responses is being used by which receiver and for what purpose. Finally, there are strong recurrent interactions between V1 and V2–V4, meaning that the (hierarchical) model of a unidirectional communication channel with a separable sender and receiver breaks down.

If we move forward to areas like IT (or a deep layer of a neural network), then one could say that the "neural code" (Type 1) becomes increasingly more "usable" higher up in the processing hierarchy, in the sense that it becomes easier to do something meaningful with it (e.g., a face-selective IT neuron or hippocampal place cell), although there's inevitable data loss compared to lower areas (data-processing inequality). This process can be thought of as a series of "unfolding" transformations that create increasingly linearly separable manifolds, corresponding to object categories, in high-dimensional spaces (e.g., Chung et al. 2015). The quantification of how "usable" and efficient a code is might be critical to interpret the significance of mutual information quantifications between stimuli and neural responses. For instance, we can, in principle, decode object categories more accurately from the retina than from area IT (data-processing inequality). This does not mean that the receiver of the retinal output uses this information to make decisions about object categories and act upon them. However, information about object categories, and the image in general, can be easily decoded from activity in area IT and is highly compressed.

Coding in the Motor System

Turning toward the motor system, it becomes apparent that the notion of a communication theoretical code (Type 1) becomes problematic for many reasons. Information about many different movement parameters is present in the firing rates of motor cortical neurons. This information is encoded in the motor cortex. The problems begin with *how* the information was encoded: Did the encoding occur before arriving in the motor cortex, in local circuitry, or as input to the particular neuron, whose action potentials are being recorded? The time-varying values of different movement parameters tend to be correlated, reflecting the complex mechanics of movement where many degrees of freedom vary simultaneously. M1 neurons have high-dimensional tuning curves, so that the firing rates of individual neurons contain information about many of these parameters. This makes it difficult to parcel the encoded information into separate categories. Although the motor cortex is often considered to be composed of upper motoneurons projecting to the "final common pathway" (Sherrington 1906), in reality, it is one of many inputs to the subcortical neuronal substrate of muscle contraction. This presumed role of the motor cortex in muscle contraction has fostered historical controversies pertinent to the idea of whether decoding takes place at all in the projection targets of these neurons. As an extreme example, if motor cortical neurons function merely as upper motoneurons, then the information contained in their firing rates does not need to be decoded at all, since the putative role of these neurons is solely muscle activation. In contrast, if the encoded information is pertinent to more cognitive issues, such as the intended action of a hand on an object, then for this information to be realized as behavior output, "decoding" must take place as it is transformed by "downstream" structures to "cause" muscle contraction.

Further examination of implicit assumptions might help focus these issues. There is a general tendency in neuroscience to view the nervous system as discrete, separate components. This stems from historic anatomical descriptions of the system as well as the clinical observations underlying neurology and neuropsychology, which focus on finding localized lesions in the system. Furthermore, since the industrial age, we have become comfortable with the idea of machines composed of individual parts, each with a specific function. These factors come together to reinforce the general simplistic notion of cause and effect that underlie most functional descriptions of nervous system operation. Structure A projects to Structure B, contributing excitatory input to B's neurons, and these are diagrammed as sticks with plus and minus signs between boxes for each structure. Of course, these "circuit" diagrams rapidly increase in connections as more results are added, but the boundaries between the boxes remain fixed even as the number of sticks increases. Although it is obvious that many inputs interact to "cause" an output, such consideration is usually set aside to "simplify" neural functioning, keeping functional description within the bounds of simple causality.

This predilection toward simplistic causal circuitry is manifest in classic visual system neurophysiology. Here the concept of hierarchical organization prevails. Processing starts in the retina where coding begins, and rods and cones pixelate the visual scene. The pixel information is then transmitted to subcortical and cortical structures. As this information traverses successive brain structures, it is transformed successively. The concept here is that visual information is molded into a coherent image, one that is ultimately realized as a perceived, accurate description of the world. This concept originated with Hubel and Wiesel. They found that neurons in the cat thalamus and visual cortex had receptive fields of various complexity and hypothesized that increased complexity resulted from successive stages of processing. This concept prevailed in ensuing years during which researchers found that neurons in cortical areas anatomically farther from V1 seemed to have response properties that encompassed a wider set of visual filters. This was the motivation for attempts to organize the multitude of vision-related cortical structures into a coherent framework. Van Essen and colleagues developed a set of anatomical criteria to delineate different vision-related structures and to categorize the anatomical connections between them (Felleman and Van Essen 1991). Of particular relevance here was the idea that projections originating only from superficial cortical layers and terminating in layer 4 of the target area transmitted information in the *forward* direction, whereas those coming from both deep and superficial layers terminating outside layer 4 were receiving *feedback* information. In this case, *forward* means ascending the hierarchy with feedback in the opposite direction. Felleman and Van Essen (1991) considered the difficulty of resolving reciprocal and lateral connectivity into the scheme and suggested that hierarchical structure could exist even without stepwise serial processing. For this reason, they extended the basic anatomical criteria and added a third category of lateral connectivity to build the canonical Felleman–Van Essen diagram. This scheme consists of boxes, corresponding to specific structures,

From "The Neocortex," edited by W. Singer, T. J. Sejnowski and P. Rakic.
Strüngmann Forum Reports, vol. 27, J. R. Lupp, series editor.
Cambridge, MA: MIT Press. ISBN 978-0-262-04324-3

vertically arranged into hierarchical levels. The arrows connecting the boxes are based on anatomical tracing data.

Although this notion of hierarchy was inferred from neurophysiological experiments of function, in the Felleman–Van Essen diagram, only anatomical criteria were used. From a functional standpoint, reciprocal connectivity is not easily resolved into a flow of information. In terms of causation, the relative timing of discharge between interconnected sites might be indicative of transmission direction. As has become apparent from cross-correlation studies (Moore et al. 1966; Perkel et al. 1967; Gerstein and Perkel 1972), however, simple causal interaction between pairs of neurons is very rare. This issue is exacerbated with box-and-arrow diagrams, suggesting that information is processed in successive hierarchical levels with well-defined borders, implying that information enters as discrete input and leaves as transformed output with the complete operation taking place within the confines of the structures comprising that level. This logic is engrained in theories of sensory processing. In terms of encoding and decoding, this theme would suggest that input to a processing stage would need to be decoded and then encoded as output transmitted to the next stage.

This conceptual framework is difficult to apply to motor systems. Continuing the hierarchical logic, the general inference is that raw sensory input is processed successively to form a consciously perceived percept of the world. This takes place in well-defined anatomical structures, and according to the Felleman–Van Essen diagram, the hippocampus is the pinnacle where the percept crystallizes. From there, other cortical operations take place leading to a well-formed decision to achieve a particular goal. The goal is then transmitted to the motor system to produce the movement that achieves that goal. However, to find evidence for this scheme, it is necessary to identify the input to the system. Support for this type of post-decision signaling has proved elusive. Furthermore, many different anatomical structures project to the motor system and these projections do not follow a successive sequence of clearly defined serial processing steps.

Similar problems underlie the controversies of whether the primary motor cortex (M1) functions directly and primarily to generate muscle contraction or, instead, in the formulation of higher-level behavioral planning that gets transformed to muscle contraction as it "descends" a hierarchical structure to spinal motoneurons. Anatomical evidence shows that a small component of M1 output projects directly to spinal motoneurons and historic electrical stimulation of M1 results in somatotopic muscle contraction, which would support the idea that M1 functions to contract muscles. The counterargument is supported by recording experiments that extract movement information related to the velocity of the arm, wrist, and fingers during movement (Wodlinger et al. 2015). In the reverse hierarchical scheme, this would be "downstream" from muscle contraction in terms of execution (muscle contraction "causes" limb displacement), but "upstream" when considered as a plan (muscles are contracted to make the arm move according to a plan). Since this information appears in the motor cortex well before muscle contraction, this could support the argument that M1 functions in high-level movement planning. It should be noted that signals reflecting muscle EMG can also be extracted from M1 activity (Humphrey 1986; Townsend et al. 2006; Pohlmeyer et al. 2007) which further clouds this argument. At this point it is useful to reinforce the distinction between coding and code. The ability to extract muscle or movement information from M1 activity shows that this information exists and has been encoded (somewhere). What it might be used for (i.e., where it is being decoded) is a separate issue. As for the information content, there are at least two explanations for how muscle and movement information can be extracted from the same population of neurons. First, the disparate information may be encoded in a high-dimensional space, as seems to be the case for at least ten different kinematic parameters (Wodlinger et al. 2015). In this case, the muscle and kinematic parameters would simply occupy different dimensions. Second, the muscle and kinematic parameters are correlated (Todorov 2000; Reina et al. 2001; Scott 2003). This would suggest that both parameter sets share a single input source and that M1 activity is also related to that source. A subsequent "decoding" stage that separates these parameters may not even be needed, if the common muscle-kinematic signaling is formatted to contribute to muscle excitability.

The idea of hierarchy comes into play again in these issues. Area M1 may not be a singular node where information converges in an exclusive sense; this convergence may occur only in the executed movement. Instead, information about movement may be highly distributed throughout many interconnected structures of the motor system (and probably other parts), making it difficult (and perhaps improper) to designate a neural signal as an input or output. Since synaptic integration is a fundamental property of nervous systems, and in mammals there are typically thousands of converging dendritic inputs and as many diverging axonal terminals, the ability of any single or small group of synapses to "cause" a downstream event is small. This means that simple cause-and-effect arguments have limited utility in explaining function. It is important to consider the nervous system in its actual complexity and to realize that conventional concepts of discrete circuitry based on straightforward causal logic has placed severe limits on our understanding of the nervous system.

Conclusion

In practice, for the vast majority of neuroscience studies, we are still at the stage of figuring out what information neuronal populations contain on longer timescales; the many unknowns stipulated (e.g., information on short timescales, goals of encoding, relevant receivers, design principles, which information is actually being used, distributed representations, assumptions about hierarchy) imply that by and large we do not know what the neural code (Type 1) is, and how useful the communication channel model will prove to be for different systems. Other models that have been successfully able to model neuronal responses, like deep neural networks, do not have any inherent notion of coding in the communication theoretical sense, although coding concepts have been used to improve our understanding of what these networks actually do (Shwartz-Ziv and Tishby 2017). If progress is to be made, future efforts will need to go beyond quantifying what information is contained, to quantifying what information is actually being transmitted to whom and for what use.

Generalization to New Conditions and Failures Thereof

Encoding/decoding models, such as those described above, may fit the data they were trained on, but might not necessarily generalize; that is, they may not predict responses to new stimuli that have not been previously tested. We illustrate this point with an important *failure* of such generalization, using a situation which, according to the textbook understanding of vision as a feedforward representation of aspects of the retinal image, should not occur.

Since the pioneering work of Hartline, Kuffler, Hubel, Wiesel, and others (see Spillmann 2014), the notion that visual neurons have receptive fields anchored to particular positions on the retina has been a fundamental concept underpinning of visual neuroscience. Thus, measurement of the receptive field's position under one set of conditions might be expected to generalize accurately to the position of the receptive field tested under other conditions.

In structures such as V4, FEF, and parietal cortex, however, this generalization has not held. When the eyes move or maintain fixation at different orbital positions, the retinal location of the receptive field can shift to novel positions. The new receptive fields appear on varying timescales and may be either transiently present in conjunction with a change in eye position or exist stably for the duration of an epoch of fixation.

This finding has important implications for what needs to be included in models of encoding of visual information and suggests that the "label" on the line for such neurons is not an exact match to single particular retinal or eye-centered locations. Instead, eye position/movement is one aspect of the full "context vector" that needs to be incorporated into predicting how a neuron will respond under novel circumstances. Other factors in that context vector include attentional state, arousal, task context, recent stimulus history, stimuli from other modalities, and no doubt many as yet unexplored sensory and cognitive factors. We describe these variables here using human intuitive phrases, but ultimately they must be instantiated by aspects of neural architecture and neural firing, for which we do not yet have intuitive access (e.g., "n_g" in Figure 13.1).

Other examples come from earlier work in the visual system, in V1, where what is encoded depends critically on stimulus configuration. This challenges

From "The Neocortex," edited by W. Singer, T. J. Sejnowski and P. Rakic.
Strüngmann Forum Reports, vol. 27, J. R. Lupp, series editor.
Cambridge, MA: MIT Press. ISBN 978-0-262-04324-3

the concept of receptive fields: a simple cell, for example, will have unpredictable responses when challenged with complex scenes. Since the space of possible contextual modifications is close to infinite, there is no canonical definition of a receptive field. The same problem will hold for representations in general: they will change as a function of the content to be represented (encoded). This inability to establish 1:1 mapping will also pose problems for the analysis of the relation between a code and the respective behavioral consequences (as in decoding models, see Figure 13.1). We expect these challenges for models that link neural activity to behavior (i.e., decoding models) to be most severe at intermediate levels of processing, but to diminish as one builds decoding models that take as their input neural responses that are closer to the motor effectors (muscles) (see “b(t)” in Figure 13.1).

On the whole, we cannot at present assume that assessments of visual coding at the individual neuron level measured in one task will necessarily generalize to another.

Reliability, Stability or Generalizations to Repetitions of the Same Conditions: Inferring the Stimulus from the Activity

Another widely recognized problem is that even repeating the same conditions does not produce the same activity pattern. This variability in neural firing is often referred to as noise, but it is increasingly understood that what appears as noise to the experimenter is not necessarily noise to the brain but could reflect signals related to aspects of the environment or state of the organism that are not under experimental control.

Put another way, this variability means that one might not be able to reliably predict the firing pattern of the population from the stimulus. Another way of asking the question is whether the stimulus can nevertheless be inferred from the neural activity, despite this variability. Judging the type of information present in a neural population in this manner provides insight into what knowledge an organism has access to.

Reliability, in general, is defined as an invariance of a classification or identification of a state in the presence of some kind of perturbation. This perturbation can result in a change of the code or representation, as a consequence of noise or unknown states of the system. Reliability of a representation characterizes the ability to identify the encoded information from noise-perturbed observation.

In contrast to reliability, *stability* refers to a change of the representation over time. Often a stable code is understood as a constant encoding model. Additionally, stability has been defined as an error correction property that reduces the noise of a perturbed system. This definition has often been used for dynamical systems that show dynamics governed by attractors of some kind.

What Signals Constitute the Code or Are Relevant for Information Transmission?

That spikes are central elements of information transmission supporting moment-to-moment behavior seems beyond dispute. In recent decades, it has also become clear that there are temporal aspects to neural response patterns, and these temporal aspects can have powerful implications for information transmission between ensembles of neurons or between brain regions. The functional importance of spike timing has been explored in the following contexts:

- The transmission of information between neuronal populations.
- The encoding of information through relative timing among neurons, or to some external event.
- The formation of memories through spike timing-dependent plasticity (STDP) (see also Singer, this volume).

Synchronization can modulate information transmission through enhanced integration of EPSPs, as well as through dendritic nonlinearities (Salinas and Sejnowski 2001). Furthermore, synchronous volleys of excitatory inputs may effectively escape from feedforward inhibition (Fries 2015). Coherence between sending and receiving neuronal populations may bias information transmission by aligning the arrival of input spikes with windows of opportunity in the receiver (Fries 2005). One view is that information is encoded through population rate coding, but that the transmission of information is modulated by synchrony and coherence among neuronal populations, according to cognitive demands (Fries 2005). Support for selective information transmission according to cognitive demands comes from the finding that attention selectively and strongly modulates the inter-areal coherence in the gamma-frequency range, between areas V1 and V4 (Bosman et al. 2012). There is, however, ample evidence for encoding of information through spike timing. For instance, hippocampal CA1 place fields carry place information both through rate changes as well as through the spike phase relative to ongoing theta oscillations (Huxter et al. 2008). A similar phenomenon, in the gamma-frequency range, has been demonstrated in visual and frontal cortex (König et al. 1995; Siegel et al. 2009; Vinck et al. 2010; Havenith et al. 2011). Finally, the existence of STDP mechanisms shows that the timing of pre- and postsynaptic spikes critically governs synaptic plasticity formation (Markram et al. 1997; Sejnowski and Paulsen 2006). In hippocampus, neural activity shows extremely synchronous behavior during sharp-wave ripple complex, with sequential activation patterns mimicking the sequential activation of neurons during spatial navigation. These patterns are thought to be important for the consolidation of episodic memories, and interruption of hippocampal activity during sharp-wave ripples impairs spatial memory formation (Girardeau et al. 2009) as well as place field stability (Roux et al. 2017).

Relevant temporal dynamics can also be non-oscillatory. Examples of complex but not necessarily oscillatory dynamics include insect olfactory system (Wehr and Laurent 1996), leech motor decision making (Briggman et al. 2005), rodent hippocampal system/replay (O'Keefe and Recce 1993), birdsong motor system (Hahnloser et al. 2002), and primate motor cortex (Churchland et al. 2012; Mante et al. 2013; Suway et al. 2018). The concept of an "oscillation" in general suggests a static and clock-like behavior that does little justice to the nonstationary nature of neural activity (Burns et al. 2011; Xing et al. 2012) as well as to the spatiotemporal dynamics from which these "oscillations" are often the result. What appears to be an oscillation in recordings on single electrodes is often a traveling wave on arrays of electrodes (Gelperin and Tank 1990; Kleinfeld et al. 1994; Tank et al. 1994; Lubenov and Siapas 2009; Muller et al. 2016). However, as discussed by Singer (this volume), both empirical studies and simulation experiments indicate that the nonstationary and transient features of oscillations are actually advantageous for information processing and dynamic routing of neuronal activity. The underlying spiking activity is sparse, in contrast to the dense traveling waves in epileptiform activity (Muller et al. 2018).

Codes, Constancies, and Control of Behavior

The behavioral responses evoked by a sensory stimulus may be relatively rapid, simple, and stimulus-locked, such as an eye movement bringing the fovea to bear on a visual stimulus of interest. Alternatively, they can be slower and the consequence of an extended period of internal and covert processing involving a multitude of factors, such as a real-world decision to attend one university over another.

Even for comparatively simple behaviors that lend themselves to laboratory study, there is likely redundancy in the code and degeneracy in the relationship between activity patterns and behavioral outcomes. For instance, there are many different ways to achieve the same action on the environment. Consider an arm that has 4 degrees of freedom (DOF) from the shoulder to the wrist: to reach in 3D space, there are more DOFs than movement dimensions. This means there is an infinite combination of DOFs that will achieve the same endpoint movement. However, psychophysics shows us that we tend to use the same combinations (approximately) in a reliable way.

Certain DOFs tend to be linked or correlated during movement. Why this happens is not always due to mechanical or anatomical constraints as some can be violated volitionally. These invariants reflect a "choice" made by the system. For motor control, these choices seem to reflect control efficiency, minimizing the amount of information that needs to be transmitted to accomplish a goal. This general concept is usually attributed to Nikolai Bernstein (1967), who studied the structure of movement using an early form of video

From "The Neocortex," edited by W. Singer, T. J. Sejnowski and P. Rakic.
Strüngmann Forum Reports, vol. 27, J. R. Lupp, series editor.
Cambridge, MA: MIT Press. ISBN 978-0-262-04324-3

motion tracking. Bernstein articulated the concept of "motor equivalence" in which the same movement could be produced in many different ways. He used drawing movements as a prime example and emphasized the difference between metrics and topology. His examination of repeated drawings showed that the metrics of the drawn object varied between repetitions, but that the shape of the object (topology) as drawn by the same individual was constant. Furthermore, if that object was drawn on a table top or blackboard, the personal topological features remained consistent. This was also true if the object was drawn with the dominant or nondominant hand. Although the set of effectors (muscles and joints) varied greatly, the behavioral outcome (the drawn object) was the same. Bernstein then used these findings to discuss locationism in the nervous system. Since topology was invariant, he argued that this was the dominant organizing principle of motor function. Thus, topological features of the movement would be expected to have an anatomical constancy instead of muscles. He proposed a thought experiment in which neural activity could be observed in the brain. If muscles were localized in the brain, then there would be complicated zig-zag patterns of activity across the cortex because muscular activity is highly variable between movements. If that was the case, he asked, what would be the advantage of having neurons spatially segregated according to the muscle they activate?

Indeed, experimental results show that extracting movement trajectories of the arm and hand from motor cortical activity during reaching and drawing is straightforward and robust (Georgopoulos et al. 1986; Schwartz 1994). Population decoding of these movements is the basis for neural prosthetics. In contrast, extraction of the muscle activity taking place as the arm moves freely through space has proven to be much more difficult. Such generalized motor representations bear a resemblance to constancies that are familiar in sensory processing, such as our ability to assess color as a comparatively stable object property despite variation in the wavelengths that reach our eyes under different illumination conditions, or the perception that the world is not moving despite massive shifts in the retinal image with every eye movement.

The Single-Neuron Tuning Curve: A Motivating Idea Whose Time Has Passed?

Single-unit neurophysiology has, over the past four decades, focused a great deal of effort on describing the responses of each recorded neuron to a set (typically ~20) of experimental conditions chosen at evenly (typically) spaced intervals along a single, predetermined physical dimension (typically inspired by pilot studies or by prior work). Classic examples include:

- Responses of V1 to the orientation of a visually presented light bar (e.g., Hubel and Wiesel 1962)

From "The Neocortex," edited by W. Singer, T. J. Sejnowski and P. Rakic.
Strüngmann Forum Reports, vol. 27, J. R. Lupp, series editor.
Cambridge, MA: MIT Press. ISBN 978-0-262-04324-3

- Responses of M1 cells during the performance of in-plane center-out arm movements (e.g., Georgopoulos et al. 1982)
- Responses of a visual area MT cell to the in-plane direction of visual motion
- Responses of "face neurons" in monkey inferior temporal cortex

In each case, a "tuning curve" (or tuning function) is determined by fitting (e.g., least squares error fit) the neural responses (dependent variable) with a smooth, low-parameter mathematical function of the value of prespecified experimental axis (independent variable). The mathematical function chosen by the experimenter for fitting typically has a single peak over the domain of the independent variable, which is taken to be the value of that variable that is predicted to give the maximum response for that unit (the so-called, preferred orientation). In the cases of discreet experimental conditions (e.g., "face neurons"), the tuning curve is implicitly assumed to be a step function (e.g., on the X axis, the tested images can be plotted from left to right, where all images containing a face are to the right of the step up). Other parameters of the tuning curve are also typically computed and reported (e.g., the standard deviation of a Gaussian function can be taken as the orientation tuning width). In such studies, the values of these tuning curve fits are typically summarized over the entire sampled set of single neurons.

These single tuning curves have been very useful for at least three reasons: First, they demonstrate that individual neurons have response sensitivity over the measured variable (e.g., response sensitivity to the orientation of a drifting, full field visual grating). Second, because of the smoothness prior implicitly contained in the chosen mathematical tuning functions (e.g., Gaussian, cosine), they predict how individual neurons will respond to similar conditions (e.g., orientations that were not tested; images containing faces that were not tested). Third, in some cases, the tuning curve can be used as a functional marker to ask if one is recording from a particular area (e.g., strong motion direction tuning as a functional signature of area MT).

It can be argued, however, that the tuning curve has outlived its scientific usefulness, although our group did not unanimously agree on this point. We note at least three key weaknesses: First, in all sensory systems, single neurons are clearly sensitive to experimental changes along many possible axes besides the one chosen by the experimenter. This is well known and attempts are often made to compensate for this by either relegating some of this "nuisance" sensitivity to the methods (e.g., we first found the receptive field of the neuron, which is itself a tuning function over two dimensions) or handled by trying to test one or two other stimulus axes (e.g., orientation bandwidth, hue). While such attempts can be valiant, they always underestimate the complexity of the neural responses because the experimental condition space is very large (e.g., the dimensions of image space). Even more problematic, the ability of the experimenter to guess at the "best" experimental axis rapidly diminishes after even just one nonlinearity in the neural processing (e.g., V2 in vision), and appears almost completely arbitrary once one reaches very deep levels of the processing (e.g., inferior temporal cortex in vision).

This constraint is closely related to the second serious limitation: tuning curves have very limited ability to predict the responses of individual neurons beyond interpolations of the specific conditions already tested (i.e., limited ability to generalize). Thus, by definition, tuning curves do not contain generalized knowledge of the neuron's processing function (i.e., the image-computable encoding function in vision). Again, this problem gets dramatically worse the deeper one goes into the system (more nonlinearities).

A third major limitation of tuning curves is that they promulgate the idea that the goal of the field is to discover the "optimal" stimulus of each single unit, as if the single neuron is a homunculus that can offer direct insight into questions of complex human behavior. This is clearly misguided in the contemporary context of population coding, and we believe that even our contemporary ideas of population coding will look naive in another twenty years.

All three limitations are the result of the understandable desire of the experimenter (indeed, of the field) to impose a human-interpretable prior on the responses of the neuron to help organize one's thinking before reporting those responses to the world. Simply put, we would prefer it to be the case that neural responses can be reduced to a few dimensions of the domain of interest (e.g., the domain of all images) so that we can more readily communicate our findings—our "story" and our discovered "principles"—to other members of our species. As cognitive scientists, we deeply appreciate the social primate value of storytelling. But that is not the same as science that gauges its progress through accurate prediction of the phenomena of interest (e.g., neural spikes, behavior). We see no reason to assume that evolution has left us with an adult brain whose complex internals are readily communicated to other humans in such simple forms as tuning curves.

Fortunately, when we set the tuning curve aside, we do not need to go back to square one. We now have better methods of estimating much more accurately (i.e., generalize to new images) the encoding functions of individual neurons using systems identification methods combined with modern artificial neural networks that provide much more appropriate (highly nonlinear) encoding bases (e.g., Yamins et al. 2013). These methods have rapidly spread in the visual system, and somewhat in the auditory system. They have not yet been applied to all sensory systems or to motor systems, but much active work is ongoing and we expect these advances to continue apace. We also note that even these currently cutting-edge approaches will still be incomplete without incorporating models of how internal states (e.g., ongoing neural activity within the local population) predict neural responses during presentation of sensory stimuli (Dechery and MacLean 2018). Indeed, the most current encoding models for visual processing are still not able to capture the temporal dynamics of visual system neurons.

Zooming out, we also believe that the human desire for interpretability should not be forgotten but that it should be redirected. While far from guaranteed, human interpretable "principles" might still be found, but modern artificial neural network methods and experimental progress both argue that human interpretable principles might best be found at the levels of cortical architecture, learning, development, and perhaps even evolution. For example, the principles could take the form not of a set of connections or activity patterns that allow the brain to perform a computation, but of the rules of activity-dependent plasticity that enable these connections to be set up (see, Singer, this volume).

Even though the tuning curve has limitations as a research tool, it has not outlived its usefulness as a pedagogical tool. Indeed, it provides an elementary introduction to the idea of encoding functions in sensory systems (and projection functions in motor systems), which then motivates the idea of a high-dimensional, predictive response function. In this vein, tuning curves helped to promote an important conceptual advance: the idea of population coding. Specifically, tuning curves considered from a population of neurons (as outlined above) naturally suggest that, when viewed as a group, the value of the currently presented stimulus can be "reported" to downstream brain regions (population code), and they have motivated ideas and testing of how alternative population codes might estimate that value to guide behavior (e.g., examples in motion discrimination, motor control).

In sum, the idea of a tuning curve has helped carry the field toward the contemporary goal of discovering accurately predictive neural response functions (e.g., image computable encoding functions in vision) as well as toward defining contemporary concepts of population coding. However, we now know that as soon as we step beyond the very earliest stages of a sensory system, the tuning curve becomes overly simplistic as to only maintain introductory pedagogical value. Fortunately, the contemporary approaches outlined above are ready to carry the research forward.

Units of Analysis in Brain Tissue

Are Circuits Well Defined and Amenable to Study?

Considerable interest in neuroscience in recent years has focused on the concept of circuits. The general idea of a circuit comes from electronics. In that system, circuits are composed of discrete components and the connections between them are concrete. In the brain, the physical connectivity can also be fairly concrete. In some cases, different structures can also be well defined.

In the functional domain, however, this is not clear. Do defined single anatomical structures have singular functions that are different from other structures that remain constant over time? This functional idea, as expressed in typical box-and-arrow diagrams (with plus and minus signs) to describe the functional pathways of information, is wrong because it implies a high degree of discreteness that is hardwired: In a brain containing billions of neurons, we cannot define nodes this cleanly unless every neuron has its own box. More importantly, given the high degree of recurrence in brain circuits, we cannot define input and output this way.

Historically, the cortical column was viewed as a key computational element of a cortical (micro)circuit. A cortical column was initially defined (in the visual system) as a small cross section of cortex, in which neurons in the different layers of cortex share some kind of common property, beyond similarities in their receptive field position. For example, in V1, there exists a retinotopic map in two dimensions along the cortical sheet. Across layers, however, there is a similarity in the ocular dominance and orientation selectivity of neurons at a given location on that cortical sheet. There can be discontinuities in the sensitivity to orientation and ocular dominance for adjacent locations along the cortical sheet, which can be thought of as the borders of columns.

In updating the concept of a column or canonical circuit in cortex, a key observation is that there are massive numbers of excitatory recurrent connections. They are mostly local, and there is some degree of stereotypy both within and across layers (potentially also including the thalamus). It should be noted that the probability of a connection falls off exponentially as a function of distance, calling into question the idea that there are regularly repeating boundaries between circuit elements. We should thus probably think of the cortical sheet as changing in a continuous fashion, with motifs of local connection patterns repeating smoothly.

Do such motifs perform basic sets of operations that are stereotypic across cortical areas, applied to whatever the inputs of that area may be? One such operation might be a convolution using a local kernel, followed by a static nonlinearity and normalization, as employed in artificial convolutional neural networks. This analogy, in fact, viably demonstrates how powerful such a concept, in principle, is when applied to real-world pattern recognition tasks (e.g., object recognition), or when transferring it from one specific set of inputs to another (e.g., images vs. sounds).

Whether this analogy is deep or valid only on a superficial level is currently under debate. What is clear, however, is that in cortex such an operation would perform a much more complex and flexible nonlinear operation involving a number of different cell types, recurrent excitatory and inhibitory feedback (within and across different layers) and potentially employing a whole range of different temporal delays to boost computation (see below); it would also be adjustable, for example, by neuromodulation. While the instantiation of this "convolutional" operator in cortex (e.g., precise wiring diagram) might vary from site to site, the plasticity/developmental rules by which such an operator could arise may be the same across cortex (i.e., translational invariant; see also discussion of LIPU, above).

Such a concept is comparatively familiar in the visual domain, but the extension to other domains is a more complicated question. For instance, in auditory processing, widely separated frequencies that are integer multiples of a common fundamental are likely to be grouped and processed similarly but may be processed by quite distinct neural populations at early stages of the pathway. In structures like prefrontal cortex, neurons are responsive to sensory stimuli, but there is no known similarity of tuning to stimulus features in nearby neurons. It could be the case that there is some other dimension of the input space that is well ordered, but this has yet to be established.

An open question, then, is how can we leverage modern experimental and computational tools to establish the existence of such an operator and to characterize its computational capabilities? If approached anatomically, the region in cortex that we would have to analyze would likely cover several millimeters. A functional characterization would require a tight control of both the various inputs and outputs to cortex. Inputs could be assessed by calcium imaging of axon terminals that provide input from other cortical areas or thalamus. Assessing the output would require identifying the anatomical projection patterns of putative output units. Developing adequate perturbations will certainly be crucial. An interesting first step in this direction has been conducted by Constantinople and Bruno (2013), who show that silencing pharmacologically layer 4 in barrel cortex affects response properties of layer 5/6 neurons (assessed with intracellular recordings) very little, suggesting that it might be possible to study some components of the operator independently from others.

Recurrent Connections and Ongoing Activity

A key issue, which is arguably not a central element of many views of cortical coding, is the importance of recurrent connections and the elaboration of signaling in time that such recurrence necessarily involves (see Singer, this volume). Whether this recurrence is excitatory, inhibitory, or both has important implications for its impact on neural coding and function. A number of roles and effects of recurrence have been identified or hypothesized, including the preservation of information over various rather short timescales (millisecond to second range), the generation of "spontaneous" activity and activity fluctuations, as well as "handshaking" to reflect acknowledgment of signals passed from one ensemble to another, as in asynchronous computing.

As noted, not only are local neocortical and hippocampal circuitry distinguished by extensive recurrent excitatory connections, but there are also extensive long recurrent loops between cortex and thalamus, cortex and cerebellum, and cortex and basal ganglia, not to mention projections to and from attentional centers such as parietal cortex and frontal eye fields. Clearly, recurrent connections are a defining feature of neocortex. The prevalence of local recurrent connectivity has the downside of apparently making these structures predisposed

From "The Neocortex," edited by W. Singer, T. J. Sejnowski and P. Rakic.
Strüngmann Forum Reports, vol. 27, J. R. Lupp, series editor.
Cambridge, MA: MIT Press. ISBN 978-0-262-04324-3

to the pathophysiology of epilepsy, thus suggesting that recurrence must also have benefits that justify this cost.

Dark Activity

Recurrent connections give rise, at least in part, to spontaneous, ongoing activity, or activity changes that are not locked to the stimulus presentation (Figure 13.2). Such aspects of neural signaling are called "dark activity" to reflect the fact that we have very little understanding of their functional role. Early studies of reduced slice preparations demonstrated that isolated circuitry in acute slices of neocortex have a capacity for spontaneous activity. Notably

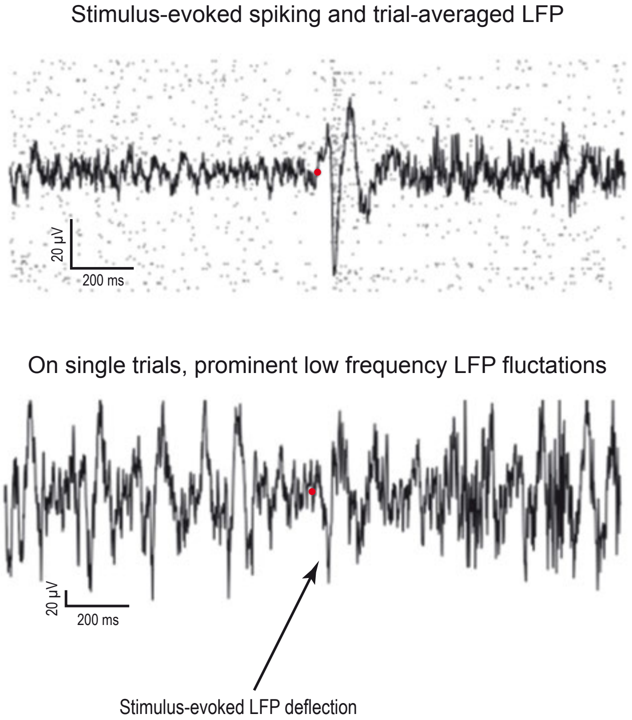

Figure 13.2 Dark activity is a prominent component of cortical activity. Upper panel: Black line shows local field potential (LFP) recorded on a single channel in marmoset middle temporal area, averaged over multiple representations of a drifting grating during passive fixation. Dots indicate spiking activity of a single middle temporal unit. Red dot indicates time of stimulus onset. The average LFP shows modest fluctuations prior to stimulus onset. Lower panel: LFP recorded on a single trial drawn from the trials that were averaged in the upper panel. LFP fluctuations are pronounced and similar in amplitude to the stimulus evoked response (arrow). Unpublished data from Z. Davis, L. Muller, T. J. Sejnowski, and J. H. Reynolds.

From "The Neocortex," edited by W. Singer, T. J. Sejnowski and P. Rakic.
Strüngmann Forum Reports, vol. 27, J. R. Lupp, series editor.
Cambridge, MA: MIT Press. ISBN 978-0-262-04324-3

it remains unclear how much this predisposition for spontaneous activity, as a consequence of local recurrent connectivity, is engaged in the ongoing or spontaneous activity *in vivo*.

Dark activity is manifested, in part, in fluctuations in the activity of individual neurons or pairwise correlations that occurs even in the absence of a stimulus. Neural fluctuations have often been treated as a form of neural noise (Zohary et al. 1994; Shadlen and Newsome 1998; Bair et al. 2001; Kohn and Smith 2005). Consistent with this interpretation, attention can reduce neuronal variability (Cohen and Maunsell 2009; Mitchell et al. 2009), possibly by regulation of lateral inhibitory circuitry (Schmitz and Duncan 2018), consistent with the notion that attention may quench noise so as to improve sensory encoding. Such fluctuations have also recently been theorized to permit preservation of multiple items in neural ensembles as a form of time division multiplexing (Caruso et al. 2018).

These forms of variability may, however, reflect other computational modes that are computed by the same neural population. It has been proposed that synchronous neural responses may act to regulate input gain (Swadlow et al. 1998; Fries et al. 2001a; Salinas and Sejnowski 2001; Sohal et al. 2009) or aid information transfer between these populations (Fries 2015). Multichannel recording approaches have revealed traveling waves of neural activity in multiple cortical areas, from sensory to motor (Muller et al. 2018). These waves can be evoked by external stimuli and can also occur spontaneously. They are likely mediated by recurrent circuits, transiently modulating neural excitability as they pass.

Moreover, ongoing activity has implications for any consideration of cortical population encoding of stimuli. It is difficult, for instance, to predict single-trial activity of any individual neuron simply knowing its tuning properties (~15% variance explained). In contrast, local population activity, including both tuned and untuned neurons, can be used to predict individual neuron activity on a moment-to-moment basis very well (up to 85%) in awake ambulating mice (Dechery and MacLean 2018). This argues strongly for a multineuronal-based coding scheme that takes into account the state of the local population, presumably dictated in large part by local recurrent connections, rather than the stimulus alone. Cortical population responses can be seen as generative rather than passive.

At a larger spatial scale, inter-areal recurrent connections strongly regulate activity. For example, within the domain of vision, recurrent circuits from the frontoparietal attentional control network, including parts of the oculomotor system (Reynolds and Chelazzi 2004), impinge on visual cortical areas. In addition to modulating ongoing activity, as noted above, these feedback signals strongly modulate stimulus-evoked responses, increasing the strength of responses to attended neurons and, through local recurrent inhibitory circuits, suppressing responses to unattended stimuli (Moran and Desimone 1985; Reynolds and Heeger 2009; Ni and Maunsell 2017).

At a more general level, recurrence of excitatory connections has the capacity to promote both inference (lateral and hierarchical) and learning. Recurrence also plays an integral part in multiple recurrent neural network models, such as Boltzmann machines, liquid state machines, and echo networks.

Activity Evolves in Time

In the past, delays were ignored in most models (especially in models that tried to identify computational principles), despite the fact that we all know that delays are omnipresent and range from sub-milliseconds to seconds. In addition, delays may result from several biological mechanisms: from conduction velocities, axonal delays, delays that result from the activation of neurons, or produced by modulatory processes in the system.

Recent models in machine learning and artificial intelligence, for instance, make use of recurrent networks to process time series, such as speech. For learning the optimal connectivity for a certain task, recurrence is unfolded over time to map time to space. For this, a constant delay between neurons and layers is assumed. Delays have also been used to implement auxiliary functions in dynamical systems. Among these is the use of delays to stabilize oscillatory dynamics and zero time lag synchronization. None of these approaches, however, addresses the effect of distribution of delays, which seem to be the most appropriate description of delays in recurrent networks.

Therefore, understanding the impact of delay distributions on computations remains a challenge. Recently a conceptual framework, *delay-coupled reservoir computing*, was introduced (Lagorce and Benosman 2015). It extends the computational principles from reservoir computing and explicitly uses single delays as well as delay distributions for the implementation of universal computation (Appeltant et al. 2011). The framework uses the concept of delay-coupled differential equations, which is a differential equation that receives a delayed and maybe transformed input back into the system.

In contrast to ordinary differential equations, this difference brings the system into a whole new category of dynamical systems that map functions onto functions (i.e., infinite dimensional mapping).

While the mathematical concept may be hard to understand in detail, there is a beautiful and simple analogical interpretation. The additional delay coupling of a single dynamical system to itself (i.e., neuron) can be interpreted as a network of virtual neurons that are recurrently connected with a constrained connectivity matrix. Thus, the combination of delay-coupled single elements into a network generates a larger recurrent network composed of the real neurons with real connectivity, and virtual neurons with virtual connectivity that are contributed by the delayed self-coupling. In other words, the effect of the delay coupling is a virtual increase in the number of neurons and an effective increase in the coding space.

In sum, the framework of delay-coupled reservoir computing helps us understand the effect of delays and distribution of delays on computation as a simple extension of the classical reservoir computing with an increased number of neurons.

Context and Network State

Another aspect of the neural code that recurrent connections may contribute to is the dependence of response properties on context, both in space and time. What is happening around the neuron at a given moment and what has happened to it previously strongly influence the neuronal response. Cortical pyramidal cells are anatomically interconnected with thousands of other excitatory and inhibitory neurons; these connections likely mediate this spatiotemporal contextual influence. On the intracellular level, these contextual influences can be observed as synaptic barrages that change the likelihood of action-potential generation, for example, by changing the membrane potential.

Considering the entire cerebral cortex at the same moment, the map of membrane potentials of all of the cortical neurons could be visualized as an excitability map. The probability of activity flowing in a particular path through the cortical network will be an interaction between this excitability map and incoming activity, which subsequently changes the excitability map. Thus, the excitability map shapes interaction networks of cortical neurons on a moment-to-moment basis, allowing a great deal of flexibility to be incorporated into cortical networks. To obtain stable perceptions and behavior, however, these highly context- and history-dependent network states are expected to exhibit stable states of activity that correspond to the stable perception or action. We propose that an important feature of the cerebral cortex is the ability to generate both stable and highly flexible patterns of activity in space and time that allow behavior to occur in both a stereotyped and flexible manner.

Cortex Cannot be Understood in Isolation

The cerebral cortex evolved in mammals, joining other more ancient structures in early vertebrates that previously supported autonomous behavior. Presumably, the cortex enhanced survival in ways that we would like to understand. Considering other parts of the brain with which the cortex interacts may help expand our understanding.

The cerebral cortex is tightly coupled to several important brain structures. The thalamus is the gateway to the cortex but it also receives cortical feedback. Interestingly, the feedback connections are more numerous but weaker than the more robust feedforward projections, with a wide range of time delays in the 10–100 ms range (Crick and Koch 1998). The basal ganglia are another partner with the cortex. The cortex projects to the striatum, which through a sequence of subcortical projections returns to the cortex through the thalamus, a loop

From "The Neocortex," edited by W. Singer, T. J. Sejnowski and P. Rakic.
Strüngmann Forum Reports, vol. 27, J. R. Lupp, series editor.
Cambridge, MA: MIT Press. ISBN 978-0-262-04324-3

that takes around 100 ms. A third loop between the cortex and the cerebellum, including the prefrontal cortex, is reciprocally connected with the lateral cerebellum. The contributions of these loops are essential for understanding what the neocortex contributes to brain function.

All of these structures are interconnected with brainstem and sensory periphery, where signals that originate in the brain can even result in changes to the sensory input. For example, pupil dilation is influenced by auditory stimuli, arousal, and likely other factors. By influencing pupil dilation, such factors can influence the light reaching the retina with consequences for subsequent visual processing. Pupil diameter is highly correlated with state of central neuromodulatory systems and the membrane potentials of cortical neurons (Reimer et al. 2016). In the auditory system, it has recently been shown that eye movements are accompanied by an eardrum oscillation (Gruters et al. 2018), again suggesting information exchange between sensory pathways can be implemented via the control of the mechanisms of transduction.

Conclusions

At the time of the Dahlem Workshop (Rakic and Singer 1988), only rudimentary knowledge was available on how cortical circuits are organized, and this information was based on the concept of a cortical column. Today we have a better idea of how the different types of neurons are connected and how they influence each other, especially the many different types of inhibitory neurons. Thirty years ago, electrodes were placed in the cortex blindly and cortical neurons were recorded whose inputs and outputs were largely unknown. Although these recordings revealed diverse response properties, the observations were correlational, and it was difficult to determine how they influenced behavior. Today, optical recording techniques have made it possible to image activity in thousands of neurons simultaneously in dense clusters, and to influence their activity with optogenetics. The emphasis has shifted from the properties of single neurons to the dynamical trajectories of neural populations in state space. Although these recordings are no longer "blind," we are still far from having a functional account of cortical processing.

We have also gone from a paucity to a plethora of computational hypotheses for how information in cortical circuits is organized. There are many ways that the features of the world encoded by cortical neurons could be combined and used to produce complex behaviors. Conceptual frameworks from information theory, Bayesian probability theory, and dynamical systems theory might all give us useful insights and predictions for experiments. Machine learning algorithms are being used to analyze the big data being generated in physiological and anatomical experiments. The convergence of deep learning architectures in artificial intelligence with cortical architectures is generating insights into how cortical hierarchies could enable object recognition in images and recognition of speech.

Over the next ten years, we anticipate that progress should accelerate rapidly, both because of improved techniques for probing and manipulating neurons, and because of more sophisticated computational hypotheses for how to interpret neural recordings. Thus, in another thirty or so years time, the participants of a future Cortex Forum should have a much better understanding of how the cerebral cortex transforms dynamic patterns of input activity, how memories are organized, and how, in concert with other brain areas, the cortex gives rise to our cognitive abilities.

From "The Neocortex," edited by W. Singer, T. J. Sejnowski and P. Rakic.
Strüngmann Forum Reports, vol. 27, J. R. Lupp, series editor.
Cambridge, MA: MIT Press. ISBN 978-0-262-04324-3