Cognition and the single neuron: How cell types construct the dynamic computations of frontal cortex
Amelia J. Christensen¹, Torben Ott¹,² and Adam Kepecs¹

Abstract
Frontal cortex is thought to underlie many advanced cognitive capacities, from self-control to long term planning. Reflecting these diverse demands, frontal neural activity is notoriously idiosyncratic, with tuning properties that are correlated with endless numbers of behavioral and task features. This menagerie of tuning has made it difficult to extract organizing principles that govern frontal neural activity. Here, we contrast two successful yet seemingly incompatible approaches that have begun to address this challenge. Inspired by the indecipherability of single-neuron tuning, the first approach casts frontal computations as dynamical trajectories traversed by arbitrary mixtures of neurons. The second approach, by contrast, attempts to explain the functional diversity of frontal activity with the biological diversity of cortical cell-types. Motivated by the recent discovery of functional clusters in frontal neurons, we propose a consilience between these population and cell-type-specific approaches to neural computations, advancing the conjecture that evolutionarily inherited cell-type constraints create the scaffold within which frontal population dynamics must operate.

Addresses
¹ Department of Neuroscience and Department of Psychiatry, Washington University in St. Louis, St. Louis, MO 63110, USA
² Bernstein Center for Computational Neuroscience Berlin, Humboldt University of Berlin, Berlin, Germany

Corresponding author: Kepecs, Adam (akepecs@wustl.edu)
Email address: ameliac@wustl.edu (A.J. Christensen), torbenott@wustl.edu (T. Ott)

Introduction
Frontal cortex is thought to underlie a wide variety of cognitive capacities—ethologically crucial functions like long-term planning [1], short-term memory, cognitive control [2], social cognition [3], and self-reflection [4]. More generally, frontal cortex acts to monitor, combine, modulate, and control activity across brain regions, on a variety of timescales, and its involvement in high-level behavior has been summarized as an additional layer of hierarchical processing within neocortex. Indeed, a wealth of studies have found that the activity of frontal cortical neurons correlates with a seemingly endless number of features—from sensory and spatial information to decisions and economic value to abstract concepts and rules. Is it possible to organize these diverse neural activity patterns into a core set of specific roles frontal cortical neurons perform in support of the computations involved in higher cognitive behaviors?

In visual cortex, a neuron’s tuning curve, that is, the correlation of neural activity with external features, is suggestive of the operation that neuron might perform. A visual cortical neuron responds only to specific visual features (like edges) in a certain region of retinotopic space, and its computational role can be described as filtering the incoming sensory data through its receptive fields [5]. Although this description is incomplete, it has inspired convolutional neural networks [6], models that not only predict the activity of visual cortical neurons [7], but have also revolutionized computer vision. Alas, the neural tuning approach has not led to the discovery of the analogous neural roles in higher cognition, or a new deep learning revolution. Rather it has just led to a taxonomy of occupants for the cortical-tuning zoo. So then, how can we begin to search for organizing principles that reveal the functions of these frontal neurons?

This question is now being tackled by new approaches, with major breakthroughs coming from two disparate directions. One approach focuses on organizing frontal neural dynamics by looking for patterns in population neural activity [8–10]. Separately, another approach focuses on classifying single-cortical-neurons into different anatomically and genetically defined cell types [11]. After reviewing recent insights from both approaches, we argue that cell-type specific connectivity motifs constrain the computational role played by individual frontal neurons.
Inferring computations from neural population dynamics

Classic approaches for interpreting neural activity rely on analysis of a subset of recorded single neuron firing rates by aligning the activity to relevant task phases and condition-averaging, to create peri-stimulus time histograms (PSTHs). While this approach has had famous success summarizing neural activity relatively close to the sensory periphery, frontal PSTHs can display baffling complexity [12]. That is, coding for task-relevant variables are randomly mixed throughout the neural population, and only by considering the contributions of many neurons together can the state of the external variable(s) be consistently read out from the neural activity [13]. This "mixed selectivity" has been observed in a range of frontal [12,14] and parietal [15] areas. Thus, facilitated by the advent of new technologies to record, and data analysis techniques to analyze large populations of neurons, considering neural populations as a unit has become the dominant approach of systems neuroscience across the brain, and particularly in frontal cortex.

As a brief background, population dynamics can be represented as trajectories in a high-dimensional space, with axes defined by the firing rate of each neuron [12]. By themselves, these high dimensional trajectories are simply a change of basis and provide no additional information over a list of classical PSTHs. However, combined with techniques to find lower-

Figure 1

a. Schematic of population analysis of neural activity, showing transformation from recorded PSTHs to neural state space, to a dimensionality reduced space. b. Schematic of changes in population activity manifolds observed in flexible context-dependent decision-making tasks. Insets show two ways (warping of manifold shape, and changing of initial condition) that population activity can be updated based on task context.
dimensional subspaces or manifolds in which neural trajectories move (collectively referred to as dimensionality reduction [16]), considering neural populations as a trajectory has inspired the application of mathematical techniques (like dynamical systems [8], geometry [17], and topology [18]) to search for patterns in neural data (Figure 1a). Population approaches have many advantages over single-neuron analyses—they often have less selection bias (because all or most recorded neurons are considered), they are applicable online and in single trials, and they have enabled brain-computer-interface based motor-control and communication for severely disabled human patients [19]. The population trajectory approach was popularized by seminal contributions in the field of motor learning [20], preparation [21], and control [22]; summarized elsewhere [9,13,23].

The neural population approach has revealed dynamical operations that are generally useful for flexible decision-making. Flexible, context-dependent sensory decision making is a prototypical case where mixed selectivity dominates single-neuron responses [12]. In context-dependent decision making tasks, as opposed to being represented by single neurons, context has been observed to change the shape of the neural activity landscape, to preferentially bias application of one decision rule over another [14,24–27] (Figure 1b). That is, if we imagine frontal cortex as a funnel into which various sensory variables are poured, and out of which comes a decision, context can change the shape of the funnel, effectively biasing which decision the trajectories roll towards. This context-dependent-change in dynamical landscape is an emerging motif often found in frontal cortical operations. Rather than changing the shape of the neural funnel, computational flexibility can also be achieved by changing the inputs or initial conditions (e.g., where the sensory-balls start out inside the funnel) of the system dynamics [28] (Figure 1). This computational strategy has been observed in interval timing reproduction tasks [29–32], where afferent and intrinsic input play the role of setting initial conditions of population dynamics.

Although we focused on studies where computation was performed through alteration of the input or shape of the neural state space, the population dynamics approach, combined with the simplifying assumption of low-dimensional single-trial latent factors [33], has also been used to predict behavioral responses during learning [34–36] and more generally on a trial-by-trial basis [34,37–42]. Further, low dimensional trajectories might be computationally desirable, as they are robust to noise [43], and may support both memory [44–46] and generalization [45].

These approaches outlined above have led to new ways of conceptualizing computations performed by neural populations, but they are not a panacea. Descriptions of population dynamics often occur in a latent space abstracted from the underlying neurons [8]. Thus population-computation hypotheses can be both hard to intuitively understand, and direct tests of these hypotheses would require arbitrary spatiotemporal control of neural dynamics [47,48]. Further, these approaches often include dimensionality reduction, a technique which can be sensitive to details like behavioral complexity, number of neurons recorded, fraction of a population encoding the variable of interest, length of recording, amount of preprocessing, and many other factors [49,50].

Carving neural representations at their joint: Identifying functional neuron types

The low-dimensional dynamics perspective is often motivated by the long-standing challenge of distilling the diverse neural tuning profiles found in frontal cortex to a principled understanding of computations. As an alternative to considering the instantaneous neural activity as a point in an N-neuron dimensional space, each neuron can be considered as a point in a space defined by its tuning to different task and behavioral features, as well as by its temporal activity profile across the trial. A handful of recent studies have examined the structure in this high dimensional space and have discovered that neuronal profiles are not uniformly/randomly distributed, rather they cluster together in dynamics and/or tuning. Under the hypothesis that neurons in these groups play similar roles in population computations, these studies provide a new link between neural identity and computational role. Thus, these studies, reviewed in this section, highlight a new perspective on cell-types, definable by functional properties (e.g., temporal dynamics, high dimensional tuning).

Analyzing neural activity during decisions that required integrating sensory and reward information, Hirokawa et al. [51] found that lateral orbitofrontal cortical (IOFC) neurons were tuned to privileged mixtures of task features which corresponded to specific “decision variables.” They developed a clustering approach to identify neurons with similar response profiles, in terms of tuning or temporal dynamics, and found that the neural population was highly structured with distinct groups of neurons having different coding profiles. Surprisingly each of the functional types had a clear interpretation based on a simple model of the choice behavior, quantitatively matching decision variables such as decision confidence, reward magnitude, and integrated value (Figure 2).

In search of functional clusters, Hocker et al. [55] reanalyzed recording data from a value-based decision task that separately manipulated reward probability and amount to understand the role of reward history under
They found multiple functional clusters of lateral OFC neurons with one cluster encoding reward history directly preceding the animal’s choice, similar to a functional cluster observed by Hirokawa et al. [51]. The authors suggest this group of neurons could play a causal role in driving reward-history behavior biases.

Namboodiri et al. [52] applied the same clustering approach to calcium imaging data, so they could investigate the activity of ventral orbital and medial OFC (vmOFC) neurons while mice learned a Pavlovian trace-conditioning task. Importantly, the different vmOFC clusters had distinct temporal (and tuning) dynamics that peaked at different phases of the task, allowing the authors to use optogenetics to specifically disrupt activity in only subsets of clusters, to tease apart differential cluster contributions to association learning, maintenance, and extinction. Finally, Onken et al. developed a statistical framework to establish the existence of categorical groupings of neurons in primate OFC [53]. Interestingly, in many cases mentioned here, the authors found a correspondence between clusters found in a space only consisting of tuning features, and one consisting of only neural dynamics. That is, neurons code for specific conjunctions of features, at specific phases in trial progression, as would be expected if the neural activity represented the dynamical evolution of a cognitive algorithm.

Together, these studies suggest that neural activity in frontal cortex might have interpretable, categorical structure along specific temporal and tuning dimensions. If this is true, why have previous studies reported randomly mixed neural activity? This discrepancy may be due to differences in brain regions or species analyzed; most studies discussed in this section assessed neural coding in orbitofrontal cortex, whereas random mixed selectivity has been reported in prefrontal cortex [23] or even non-frontal regions like posterior parietal cortex [53]. Alternatively, it is possible that what seems like random mixed selectivity might simply be a function of behavioral tasks that are either too low-dimensional, or that are ill-matched to the brain region studied [51].

To investigate how task demands might constrain neural coding strategies, Dubreuil et al. [54] analyzed recurrent neural networks trained to perform a variety of common neuroscience tasks. By analyzing their trained networks, they determined that the presence of specialized groups of neurons is necessary when task...
demands require flexible computations. Similarly, by analyzing neural dynamics in recurrent neural networks, monkeys, and in humans (using functional magnetic resonance imaging), all trained to perform a context dependent decision-making task, Flesch et al. [55] found more evidence for categorical computations than for random high-dimensional projections.

Thus to observe categorical encoding it seems critical to use appropriately cognitively demanding tasks that provide and require a rich embedding of the task-relevant variables.

We emphasize that these theoretical results do not simply apply to the task the animal is performing, but rather to the computations required of the brain region of interest. For example, in the case of Hirokawa et al., the behavioral task required that animals form a decision integrating time-varying reward expectations with sensory uncertainty, a flexible computation with multiple components that critically depends on OFC [56]. To drive home this point, as a counterexample we would not expect that neurons in visual cortex would display categorical encoding of these specific decision variables, even in an identical task.

In summary, if one wishes to adjudicate between categorical and randomly mixed coding types, it is important that the task provides a sufficient number of coding dimensions and that those dimensions fall within the brain region’s computational roles. Thus, given the important role the frontal cortex plays in flexible, context-dependent decision-making, these theoretical and experimental results support the idea that distinct cognitive operations might be realized by distinct groups of frontal neurons each with specific algorithmic roles. Can we find a correspondence between these functional neuron types and the more well-known structural neuron types?

Cell-type-specific computations in frontal cortex

The last decade has seen a revolution in our toolkit for categorizing the taxonomy of cortical neurons, based on gene expression [57], projection target [58], and single-cell morphology [59]. These advances have led to the identification of specific functional roles of different cortical inhibitory neuron types, such as a gain modulation or disinhibition [60,61]. There is also increasing appreciation for a division of labor across inhibitory neuron types during behavior with a potential to control the flow of information across regions at behavioral time scales ([62–67]). Similarly, new studies have demonstrated the organization of long-range subcortical connectivity of frontal pyramidal neurons [68–72] into motifs, often suggestive of an operational role. For example, a projection from the prefrontal cortex to the basal forebrain might be suggestive of top-down control of arousal [73].

In a follow-up experiment to the functional clustering work discussed above, Hirokawa et al. [51] used electrophysiology and optogenetic tagging to record from a subset of OFC neurons that project to the striatum (Figure 3a–c). These neurons were astonishingly uniform in their activity profile, with sustained activity persisting from reward receipt through the inter-trial interval, and only diminishing with the onset of the next trial’s stimuli. These neurons all belonged to one of two functional clusters that encoded positive and negative integrated value, respectively. Intriguingly, using a different behavioral task, Hocker et al. [74] identified a functional cluster of OFC neurons whose dynamics bore a striking resemblance to the striatum-projecting OFC neurons in Hirokawa et al. (Figure 3d). These neural responses show key features required for a temporal credit assignment mechanism critical in reinforcement learning [75]. In two entirely different tasks, Terra et al. [76] and Bari et al. [77] found the medial PFC-striatum projecting neurons also displayed sustained activity across task phases. Taken together, these results reveal that the frontal cortex to striatum projection comprises a core circuit module for prolonged communication of decision variables to the behavioral learning and execution circuitry of the striatum.

The cell-type-specific architecture of cortical dynamics has been extensively studied in a remarkable series of experiments by Svoboda and colleagues in motor cortex. Guo et al. [78] analyzed a subset of layer 5b neurons that project to the thalamus, which in turn project reciprocally back to the same local cortical region. Such recurrent loops are a ubiquitous feature in cortical architecture [71], and thalamocortical loops in particular are crucial for controlling cortical excitability — runaway activity in these loops can lead to epileptic seizures [79]. Indeed, the authors demonstrated that these thalamocortical loops in mice were responsible for sustaining delay period preparatory motor activity, in anterolateral motor cortex (ALM) during a delayed response task. Economo et al. [80] further elucidated the role of different subgroups of pyramidal projection neurons—while a subset of neurons in the upper part of layer 5b neurons project to the thalamus, a distinct subgroup of pyramidal neurons in the lower part of layer 5b project to the medulla, a motor output nuclei in the brainstem. They demonstrated that these anatomical projections are the circuit mechanism through which ALM pyramidal neurons were able to segregate preparatory motor activity and activity that drives motor output (Figure 3e–g). These results provide a cell-type-specific solution for a long-standing mystery in motor neuroscience: how does preparatory activity in motor cortex exist without driving motor output? A similar role for thalamus projecting medial PFC neurons in
sustained activity has been identified by two different groups, Bolkan et al. [81] and Schmitt et al. [82], pointing to a cell-type-specific computational motif, in which thalamocortical loops support sustained activity.

It is important to note that neural activity is not always found to be separable along axes defined by projection-output specified cell types. For example, Spellman et al. [83] find heterogeneous coding of specific task-features in PFC-striatum projection neurons (however, they did find a laminar gradient of reward coding in mouse PFC, following the distribution of input axons from the anterior cingulate cortex). Negative results in projection-specific coding studies could arise for several reasons, including the lack of a specific need for that projection in the task at hand, or the dominance of another projection which anatomically overlaps with the studied projection. Alternately neural coding in a given pathway might seem “mixed” when considered with respect to particular experimental variables of interest—however, there might still be patterns in the tuning to all task features (especially given high enough task complexity), and in their dynamical time course. This less biased search for patterns in neural activity reflects a search for an algorithmic role for neural activity (e.g., the maintenance of sustained activity, which in turn is a mechanism of working memory), rather than the correlation of neural activity with experimenter pre-supposed task features.

Thus, we refer to “functional cell-types” in cases when a group of neurons is found to play a shared computational role. Functional cell-types need not be connectivity defined or genetically distinct, however given that...
connectivity and biophysical properties determine neural dynamics, it is our belief that in most cases “functional” cell-types will overlap with those defined biologically. Further elucidation of the prevalence of functional types will be necessary to determine whether these types should be considered as computational primitives necessary for flexible cognitive behavior, as predicted by theory [54].

**Future & outlook**

Broadly speaking, our hypothesis is that cognition is constructed from building blocks of certain core dynamical computations, with distinct contributions from physiologically different types of neurons. The Lego set of cognition may include several of these cell-type-specified building blocks, as well as general learning-based modules whose activity is only understandable at a (sub)population level. We hope the studies we have reviewed here impress the reader that despite the many advantages of population approaches, there is often practical utility in anatomically grounded approaches. We see no reason that just because some neurons in a particular region or task display heterogeneous, uninterpretable single-neuron coding, that the entire program of identifying individual neurons with interpretable roles should be abandoned.

Frontal cortex, like the rest of neocortex, has stereotyped microcircuitry with specific long-range input and output connections. Theoretically, such network connectivity is known to determine individual neuron dynamics [84–86] and subcortical output targets are often tied to specific sensors and effectors (e.g., the retina, but also hormone release, etc.), strongly constraining their function. From an evolutionary perspective, it seems unparsimonious that biology would go to the trouble of substantial entropy reduction to create and genetically encode the exquisite order of brain circuit architecture [71,87], if the repeated long-range and local circuit modules did not play some adaptive role. Indeed, there is increasing understanding about evolutionary conservation of key long-range projections and other cell-types in vertebrates [88–90] that may be constrained by shared ecological objective functions [91].

We propose that biological cell-types with developmentally specified architecture and physiology can be thought of as an evolutionary prior on the type of computation performable by a given set of neurons. The mapping between computation and biological cell-types is bound to arise as evolution retains architectures that provide useful constraints on neural dynamics across ecologically relevant behaviors. As a practical consequence, observing the activity of specific cell-types can serve the same role as dimensionality reduction across anonymous neural populations—providing an interpretable handle on computations.

Our hypothesis can be seen as a marriage between the “Hopfieldian” (neural dynamics are the primary level of explanation for cognition) and the “Sherringtonian” (single neurons and their interactions are the primary explanation for cognition) views. As argued by Barack and Krakauer [92], we propose that neural dynamics are indeed the appropriate level at which to search for encoding of cognitive variables. However, unlike Barack and Krakauer, we believe these dynamics can be (and indeed must be, in cases when task complexity is high and connectivity is non-random [54,84]) implemented by functional classes of neurons, corresponding to groups of neurons with specific non-random connectivity motifs. If we are correct, the field’s increasing knowledge of cell-type-specific neural circuitry will provide badly needed prior information in the search for what computations are implemented by neural populations—an under-constrained problem in any realistic experimental regime.

Thus, rather than be disheartened by the need to include more biological detail into our models, we are excited about the opportunity those elements provide in progress towards understanding the frontal cortex and beyond. Not only will such a biologically grounded understanding of neural computations help us validate and generalize our theories, but it holds the potential for new understanding of neuropsychiatric treatments—that typically act, through drugs, at the level of neurons and the receptors they express, and only indirectly influence population dynamics.

**Conflict of interest**

None.

**Acknowledgments**

We are grateful to the National Institutes of Health for funding through R01DA038209 and R01MH097061 (AK).

**References**

Papers of particular interest, published within the period of review, have been highlighted as:

* of special interest
** of outstanding interest


The authors show that prefrontal cortex dynamics during context dependent decision-making display time-evolving dynamics, with both relevant and irrelevant sensory information maintained until choice by a multi-dimensional population representation.


The authors show that prior beliefs about interval timing duration temporally "warp" population population trajectories during decision making. This provides a population-based mechanism for Bayesian integration of evidence and prior beliefs.


The authors found that following different stimulus-response rules warped the geometry of neural representations in orbitofrontal cortex and striatum, expanding along the rule-relevant dimension and shrinking along the rule-irrelevant one.


The authors provide evidence that neural trajectories in dorsomedial frontal cortex in a timing task could be best explained as dynamical system in which task demands reconfigure the systems initial conditions and input.


The authors study the flexibility of behavioral speed control and identify a temporal invariance of firing rate dynamics that can be explain by recurrent network dynamics. Temporal scaling can emerge naturally in recurrent networks with scaling controlled by external input.


The authors use targeted photoablation to test the role of individual cortical neurons in distributed representations. They show that the selective degradation of representations similar to the ablated neurons can be explained with recurrent networks producing input-specific amplification.


The authors found behavioral generalization was supported by particular geometries of complex neural representation in hippocampus and prefrontal cortex, changing with task events but invariant to linear readout.

44. Cavanagh SE, Towers JP, Wallis JD, Hunt LT, Kennerley SW: Reconciling persistent and dynamic hypotheses of working


52. Hirokawa et al. demonstrate that neural activity recorded in rat OFC during a value-guided sensory decision-making task can be categorized into distinct functional clusters, and neural activity in each of these clusters can be related to canonical “decision variables” – components of a normative model of the animals’ behavior. Finally, the authors record from a connectivity-defined cell type, OFC neurons projecting to the striatum, and demonstrate that the recorded neurons carry to a single decision variable.


54. Hirokawa et al. demonstrate that neural activity recorded in rat OFC during a value-guided sensory decision-making task can be categorized into distinct functional clusters, and neural activity in each of these clusters can be related to canonical “decision variables” – components of a normative model of the animals’ behavior. Finally, the authors record from a connectivity-defined cell type, OFC neurons projecting to the striatum, and demonstrate that the recorded neurons carry to a single decision variable.


56. The authors identify categorical, non-mixed, representations of decision variables such as often value and choice outcome in orbitofrontal cortex during economic decisions.


58. The authors find that neural networks trained to perform flexible, context-dependent sensory decision-making tasks rely on categorical groups of neurons with different connectivity and coding patterns, while they find that networks trained on simpler tasks do not require these separate groups of neurons, and instead rely on connectivity and neural activity that appears approximately gaussian.


60. The authors analyze simple recurrent neural networks trained on context-dependent decision-making tasks. Using either low-variance, or high-variance weight initializations, they demonstrate that RNNs can solve these tasks using either a “rich” categorical type encoding, or a “lazy” high-dimensional (and mixed-selectivity) coding strategy. Using their RNN models they make distinct predictions about how population activity they expect under each respective coding strategy and demonstrate that neural activity in monkey PFC and human frontoparietal networks resembled the “rich” regime rather than the lazy regime.


76. The authors find a one-to-one mapping between transcriptional and projection neuron types in prefrontal cortex but heterogeneous task encoding. Prefrontal cortex to PAG projection neurons preferentially encode choice while contralateral projections tend to encode reward context.


81. Hocker et al. analyze data from rat OFC and find neural activity can be segregated into distinct functional clusters. They identify a cluster whose activity is sustained throughout the ITI and represents reward history. The activity of neurons in this cluster b ore a striking resem-

82. Chau BKH, et al.: Contrasting roles for orbitofrontal cortex and amygdala in credit assignment and learning in maa-

The authors show that dorsomedial prefrontal cortex neurons that project to the striatum show selectivity firing rate dynamics, persistent activity during behavioral inhibitory control. Turning off this projection impairs behavioral inhibition.


The authors find that functional responses of prefrontal projection neurons do not vary with projection target but rather follow a topographic gradient, which matches the pattern of specific inputs to PFC.


The authors study recurrent networks with minimally structured connectivity and identify testable experimental predictions for the relationship between connectivity, low-dimensional dynamics, and computations represented.


