

# 5

## Neurophysiology

### Differentiating Functional Contributions across Prefrontal Cortex

Erin L. Rich and Bruno B. Averbeck

#### Abstract

A long history of research in neuropsychology has supported the idea that there is functional specialization within the prefrontal cortex (PFC). To better understand how a region subserves a specific function, neuron activity is often recorded from multiple areas as subjects engage in prefrontal-dependent cognitive tasks. Contrary to expectations, these studies have generally found that neurons across PFC encode all manner of task-relevant information, with relatively little difference among regions. These data are important because they demonstrate the vast representational capacity and flexibility of PFC, yet they have been less useful when trying to glean a mechanistic understanding of how regions differ and interact with each other. In this chapter, these data are first reviewed, then considerations are proposed that might better direct future studies. Discussion includes the anatomy and evolutionary origins of the primate PFC, which suggest a gradient organization, with a main division between dorsal and ventral trends, rather than a series of smaller discrete regions. These gradients are observable in neural recordings within and across regions and may provide insights into the functional organization of PFC. It is important to note that gradients are consistent with functional differentiation across PFC, but they suggest continuous rather than discrete changes in function. Second, recent advances in neural analysis are reviewed, which focus on representations and temporal dynamics in neural populations, as opposed to individual neurons. These population codes may reveal unique insights into local function and cross-regional interactions and help us understand the unique properties of the main divisions of PFC.

#### Introduction

The idea that the brain can be divided into functional regions dates back to the 19<sup>th</sup> century. While functions of motor and sensory regions were quickly

discernable, there has been significantly more debate about parcellation of function in the prefrontal cortex (PFC). Early attempts to understand functions of the PFC led Penfield to believe the region was “uncommitted” at birth and specialized function was learned over a lifetime (Penfield 1965). However, investigations over the ensuing decades have supported the notion that functional specialization not only exists within PFC but is consistent across individuals and species. This is largely based on neuropsychology studies that find reproducible patterns of behavioral alterations following damage or dysfunction in subregions of the PFC. Logically, then, one would assume that the activity of neurons in a subregion should reflect its function. Indeed, there are many instances where neurophysiological correlates are found in the same region where lesions impair a particular function. For example, after finding that inactivation of the lateral PFC impaired performance in the delayed-response task, Fuster and colleagues recorded from this region to search for neural responses that underlie this dependence (Fuster and Alexander 1970). They found that neurons displayed elevated firing rates in the delay period of the task, which was interpreted as the neural mechanism maintaining information in mind to perform the delayed-response task (Fuster and Alexander 1971). This ability was later dubbed “working memory.” Since then, however, elevated delay period activity has been reported in a wide variety of brain regions, including other frontal areas, such as the frontal eye fields, orbitofrontal and medial frontal cortices (Chafee and Goldman-Rakic 1998; Enel et al. 2020; Kamiński et al. 2017), as well as nonfrontal areas including parietal cortex, inferotemporal, medial temporal, auditory, and temporal pole regions (Chafee and Goldman-Rakic 1998; Fuster and Jervey 1982; Gnadt and Andersen 1988; Kamiński et al. 2017; Kornblith et al. 2017; Nakamura and Kubota 1995; Napoli et al. 2021). Therefore, elevated delay period activity is not a unique property of regions required for working memory tasks. To complicate matters further, there have been demonstrations of intact working memory in the absence of elevated delay period activity (Lundqvist et al. 2018). Although there could be many explanations for these discrepancies, working memory stands as an example of a pattern that has played out in many subfields focused on different cognitive functions putatively localized within subregions of the PFC. Neuropsychology studies implicate functional localization and initial recording studies find logical task correlates in the corresponding brain region, but these are followed by tempered enthusiasm when it is realized that the signals are neither unique to that region nor reliably found there in different task scenarios. Overall, it is now safe to say that functional localization is less apparent in neurophysiology than anticipated. This conclusion has led to a resurgence of the idea that, while some specialization is inherent in anatomical connectivity, the dominant regime is that of distributed, homogenous functionality across PFC.

Here, we propose that we should not dispense with the idea of functional localization at the level of neurophysiology. Instead, we highlight two considerations for future studies. First, we review anatomical evidence that PFC may

be organized by gradients rather than discrete boundaries and consider how this might impact neural responses within and across regions. Gradient organization is consistent with function varying across PFC and therefore could be consistent with results from lesion studies. Most lesion results are interpreted, however, as evidence for functional localization within architectonic areas, which are circumscribed areas whose function is often thought to not depend on their two-dimensional location on the cortical sheet. Gradient organization, rather, suggests that function varies continuously across the cortical sheet with few clear areal boundaries. Second, we suggest that advances in large-scale recording and corresponding analysis techniques provide more valid measures of neural mechanisms and may ultimately help to differentiate functional regions of PFC. We limit our focus to nonhuman primates, where there is abundant neurophysiology data and reliable similarities to humans in prefrontal anatomy and function, but we note that there are a number of excellent reviews on functional organization of frontal regions in rodents (e.g., Heidbreder and Groenewegen 2003; Laubach et al. 2018).

### **Functional Localization from the View of Neuropsychology**





Our strongest framework for understanding how functions localize in PFC has come from examining the consequences of circumscribed lesions or other manipulations that create loss of function. Striking contrasts are found between the lateral regions, particularly the areas surrounding the principal sulcus, compared to the ventral and ventromedial regions. In general, damage to the lateral PFC produces deficits in processes like working memory, attention, and planning, often grouped together as cognitive control or executive function. On the other hand, lesions to the ventral frontal cortex produce disturbances of emotional processing, including emotional regulation and social behavior, primarily dependent on the ventromedial regions, as well as evaluation and decision making, primarily dependent on the orbitofrontal regions. Data on the medial PFC, including anterior cingulate cortex (ACC), are more mixed, with proposed functions including linking goals to actions, signaling or adjusting to errors, or using contextual information to interpret outcomes (Kolling et al. 2016a).

Based on this evidence, it is widely held that, although the major divisions of PFC work together to orchestrate behavior, they each contribute a unique function. There are still many open questions relating to the precise nature of these functions, as well as the anatomical locations that produce certain effects on behavior. For instance, more localized lesions can sometimes parse effects further yet at other times result in no detectable deficits where a broad manipulation did. Moreover, the lack of behavioral effect following a lesion does not definitively indicate that the lesioned area is not involved in the task. Behavioral measures commonly obtained in these studies, such as percent

correct or reaction time, are coarse and do not preclude the possibility that the contributions of the impaired neurons are simply not measurable at this level. Alternatively, another intact region may be able to compensate for the loss of neurons elsewhere, which is particularly important in the case of permanent lesion, when plasticity could take place over time. Despite these caveats, it is indisputable that reproducible patterns of behavioral alterations do occur following damage or interference to different regions of PFC, with clear parallels across species. This supports the widely accepted notion that there is, indeed, functional specialization in PFC. For further discussion on the degree and evidence for parcellation of function within frontal cortex, see Chapter 8 by Murray et al. (this volume).

Given this conclusion, one would expect neuron responses to differ across regions of PFC. In particular, neurons in different regions should be driven by, or *encode*, different factors related to ongoing behavior or cognitive processes. We use the term “encode” operationally, meaning that variance in a neuron’s activity is explained by variance in an experimentally defined feature, such as stimulus identity, direction of a motor response, or current task rule. This premise has guided the design of neurophysiology studies in the PFC for decades. A typical approach is to record from a specific region during a task that is impaired by loss of function in that region. Such experiments commonly reveal neural correlates of the task being performed. For instance, neurons in dorsolateral regions (dlPFC) encode information held in working memory (Chafee and Goldman-Rakic 1998; Constantinidis et al. 2018; Funahashi et al. 1989; Fuster and Alexander 1971; Goldman-Rakic 1995; Kubota and Niki 1971; Lara and Wallis 2014; Niki 1974; Niki and Watanabe 1976; Watanabe et al. 2006) as well as rules or categories in cognitive tasks (Blackman et al. 2016; Freedman et al. 2001, 2002; Wallis et al. 2001; White and Wise 1999), and neurons in orbitofrontal cortex (OFC) encode the value of choice options as well as expected and received rewards in decision-making tasks (Critchley and Rolls 1996; Hosokawa et al. 2005, 2007; Kimmel et al. 2020; Morrison and Salzman 2009; O’Neill and Schultz 2010; Padoa-Schioppa and Assad 2006, 2008; Padoa-Schioppa and Conen 2017; Rich et al. 2018; Setogawa et al. 2019; Tremblay and Schultz 1999, 2000).

Although these results are consistent with neuropsychology data, further investigation has revealed a more complicated picture. If we expect that computations differ across regions of the PFC, and our behavioral tasks can uniquely tax these abilities, then this leads to a few concrete predictions, illustrated for OFC and dlPFC in Figure 5.1. First, if a task is impaired by inactivation of region A and not region B, then neurons in region A should carry more task-relevant information than region B, or at the very least, neural responses in the two regions should differ measurably (columns of Figure 5.1). Second, if a region is required for task X and not Y, then neurons in this region should encode more task-relevant information during task X than task Y, or at least they should differ measurably (rows of Figure 5.1). Dissociations of this sort have

	Value-based task	Cognitive control task
<b>Effects of lesion</b>		
OFC 	<b>X</b> <i>impaired</i>	<b>✓</b> <i>intact</i>
dIPFC 	<b>✓</b> <i>intact</i>	<b>X</b> <i>impaired</i>
<b>Expected neurophysiology</b>		
OFC 	<b>↑</b> <i>more active or selective neurons</i>	<b>=</b> <i>baseline or less active neurons</i>
dIPFC 	<b>=</b> <i>baseline or less active neurons</i>	<b>↑</b> <i>more active or selective neurons</i>

**Figure 5.1** Conceptual comparisons between orbitofrontal cortex (OFC) and dorso-lateral prefrontal cortex (dlPFC), from the perspective of neuropsychology and neurophysiology. The top half shows the general framework supported by loss of function studies, where OFC is important for performing value-based tasks and dlPFC is important for various cognitive control tasks. This leads to the prediction that neurophysiology should vary across regions and tasks in a similar manner (bottom half).

been sought in many instances to better understand the unique contributions of different regions of PFC. However, one of the striking outcomes has been that differences are much more limited than one might expect. Moreover, we see this equivocal outcome in published studies where there may be a bias toward identifying and reporting differences that align with each region’s presumed function; this suggests that there could be a number of unpublished observations that are even more mixed. Below we briefly outline some of these data, summarized in Table 5.1, that have led to this impression. We emphasize comparisons of ventral and lateral PFC, primarily OFC and dlPFC as an example case, because there is strong neuropsychology data to support their unique and dissociable functions.

### Contrasting Neuron Responses in OFC and dlPFC

The OFC and neighboring regions are important for emotional appraisals as they relate to decision making. However, decision-relevant information such as expected values are also strongly encoded by neurons in dlPFC (Leon and Shadlen 1999; Roesch and Olson 2003; Tsutsui et al. 2016b; Watanabe 1996), as well as supplementary and premotor regions of the dorsal and lateral frontal

**Table 5.1** Percent of OFC or dlPFC neurons encoding task variables in value-based or cognitive control paradigms, from studies that recorded neurons in both regions in the same experiment. The most consistent difference is a tendency for more encoding of spatial information, such as location or response direction, in dlPFC compared to OFC. Proportions shown are percentage of all neurons recorded in an area; \*percentages estimated from published figures.

Value-Related Tasks			
Task Variable Encoded	% OFC Neurons	% dlPFC Neurons	Reference
Any decision variable	56%	49%	Kennerley et al. (2009a)
Main effect of expected reward (by trial epoch)	5, 9, 5, 6%	7, 4, 2, 7%	Wallis and Miller (2003b)
Reward × Picture (by trial epoch)	10, 5, 8, 12%	7, 8, 9, 2%	
Reward × Location (by trial epoch)	5, 7, 4, 11%	7, 16, 11, 17%	
Stimulus	34.9%	rostral 46.0 mid 46.3 caudal 55.9%	Tang et al. (2022a)
Outcome	32.2%	rostral 39.5 mid 46.1 caudal 57.0%	
Actual payoff (i.e., reward)	45.3%	41.2%	Abe and Lee (2011)
Hypothetical payoff	16.9%	21.4%	
Juice type (by trial epoch)	13, 16, 21, 18%	10, 10, 11, 6%	Lara et al. (2009)
Receipt of reward	27%	32%	Kennerley and Wallis (2009b)
Response direction	6%	13%	
Probability of receiving reward	12%	8%	
Chosen (integrated) value	9%	14%	Hosokawa et al. (2013)
Decision type (category)	57%	68%	
Cue value	57.5%*	42%*	Hunt et al. (2018)
Action (right/left) value	6%*	17%*	
Attribute (magnitude/probability) value	23.5%*	12%*	
Spatial (location) value	7%*	16.5%*	
Cognitive Control Tasks			
Strategy	14%	12%	Tsujimoto et al. (2011)
Task rules (pre-cue epoch)	17%	29%	Yamada et al. (2010)
Abstract rules	32%	49%	Wallis et al. (2001)
Category	28%	8%	Tsutsui et al. (2016a)
Rule	26%	28%	
Contingency	48%	41%	
Strategy	12%	8%	Fascianelli et al. (2020)
Directionally-selective delay period activity	3.9%	29.9%	Ichihara-Takeda and Funahashi (2007)

cortex (Roesch and Olson 2003). These areas encode value during the delay period of working memory tasks (Leon and Shadlen 1999; Roesch and Olson 2003; Watanabe 1996), when dlPFC is believed to hold relevant cognitive information online, as well as in value-based decision making tasks (Cai and Padoa-Schioppa 2014). For example, in a task where monkeys had to weigh an amount of juice against either the delay or effort needed to obtain it, decision-relevant values were encoded by similar proportions of OFC and dlPFC neurons with only minor differences between regions (Kennerley et al. 2009). In this case, more dlPFC neurons encoded movement direction, consistent with the common finding that directional or spatial information is preferentially represented in lateral regions (Grattan and Glimcher 2014; Hunt et al. 2015; Kennerley and Wallis 2009a; Tang et al. 2022a; Wallis and Miller 2003b). Others, however, have emphasized that, although it is not as strongly encoded, spatial information is not absent from OFC (Strait et al. 2016; Yoo et al. 2018). Beyond spatial selectivity, there was very little that distinguished these regions in how they encoded decision-related information.

On the other hand, processing cognitive information, particularly rules and strategies that guide behavior, is believed to be the domain of dlPFC, yet OFC and ventrolateral PFC also robustly encode task rules (Fascianelli et al. 2020; Wallis et al. 2001; Yamada et al. 2010). In a variant of the classic Wisconsin Card Sorting Task adapted for monkeys, OFC encoded both abstract rules that define the relevant feature domain (e.g., shape or color), as well as concrete rules indicating the currently correct feature (e.g., choose red) (Sleezer et al. 2016). In addition, both OFC and dlPFC neurons encoded response strategies in a stay versus shift task, and OFC even encoded the strategy earlier (Tsujimoto et al. 2011). Taken together, the encoding properties of individual neurons tend to be primarily informed by the task the monkey is engaged in, rather than the prefrontal region where they were recorded.

It is less common to evaluate the same neurons in multiple tasks, in part because this involves training monkeys to perform tasks in interleaved fashion. Some blocked designs have been used and suggest that prefrontal neurons flexibly adapt to encode information about the current task, but do so fairly uniformly, without one particular region being uniquely engaged by one task. For instance, monkeys learned to select rewarding actions or objects in different trial blocks while large populations of neurons were recorded from the full rostro-caudal extent of principal sulcus. Across this region, neuron activity shifted between encoding the rewarded actions or objects, depending on which was relevant in the current trial block (Tang et al. 2021). Another study recorded from OFC and ACC while monkeys similarly chose a rewarding cue or rewarding action (Luk and Wallis 2013). In this case, slight differences were found in the choice phase of the task, where more ACC neurons encoded actions (16% versus 10% in OFC) and more OFC neurons encoded stimuli (20% versus 10% in ACC), but this occurred while actions, stimuli, and their associated outcomes were encoded in similar proportions during all other task

phases. Again, although small degrees of difference can be found, there is an overwhelming pattern of similarity across tasks. Finally, a recent study approached this question by recording different populations of dIPFC neurons across days from the same monkeys as they performed four different tasks, only one of which was impaired by dIPFC lesions (Tremblay et al. 2023). In this case, no metrics of neuron responses were found to distinguish the tasks. Although the expectation that there would be measurable differences is as reasonable as the expectation that two regions should differentially encode information in a given task, the supporting evidence remains quite weak.

### **Reconciling Neuropsychology with Neurophysiology**

The contradictions between lesion effects and neurophysiology data have led to different interpretations. To start, the tasks used to study prefrontal function are typically relatively simple, and it has been suggested that tasks with more complexity, that are designed to better tax prefrontal function, or those with better construct validity might find distinctions that are not found with simpler tasks. While this may be true and task design is of critical importance, it is often the case that uniform neural responses are found in tasks that are the same or highly related to those in which neuropsychology studies have demonstrated functional dissociations. This argues against the notion that more refined tasks are likely to reveal neurophysiological differences among prefrontal areas. Conversely, the impacts of prefrontal damage on human behavior are most evident in daily activities rather than highly structured laboratory tasks, suggesting that less constrained tasks may be better at tapping into the unique functions of different prefrontal regions. While this is an attractive hypothesis, there are a host of challenges in parsing and interpreting unconstrained behavior and concomitant neurophysiology. Advances in markerless-tracking algorithms, such as DeepLabCut (Mathis et al. 2018), have improved our ability to parse unconstrained behavior at the level of motor movements. Still, critical gaps between observable motor output and underlying cognitive processes have so far limited the degree to which computer vision tools have improved our understanding of PFC.

Another view notes that the dense interconnectivity of prefrontal subregions could suggest that information spreads easily, and this makes neuron responses relatively uniform. If this is the case, temporal analyses, such as latency to encode information, could reveal an origin and direction of spread, and in this way point toward specialization. For instance, similar proportions of neurons in OFC and dIPFC encode expected rewards, but encoding among OFC neurons begins about 80 ms earlier, which has been taken to suggest that reward information enters PFC via OFC and is then passed to dIPFC to influence behavior (Wallis and Miller 2003b). While this may be true, it does not explain why these signals are present in both areas. For instance, if



dIPFC represents reward values because it is a node on the path to expressing reward-guided behavior as motor output, then dIPFC lesions should produce measurable changes in motivated behaviors such as value-based decision making. Alternatively, these signals could be only passively present. However, they are curiously prevalent, and potentially metabolically costly, to be just a by-product. A related idea suggests that information becomes more shared across regions as a result of extensive practice or training, which is common in monkey studies, though this encounters the same problem in explaining why this is an efficient way the brain would operate.

An alternative proposal is the “content differentiation” model, in which different regions of PFC perform the same basic computations, but do so on different types of information, which depend on the anatomical inputs that they receive (Goldman-Rakic 1987; Zald 2007). From this view, neuron responses in different regions might appear similar because specialization arises from the large-scale circuits in which each region participates. While this is plausible, anatomical evidence has also been used to argue the opposite; namely, that different regions are specialized for fundamentally different computations, such as holding information online in working memory (Petrides 1994). Anatomically, lateral and orbital prefrontal regions differ in their cellular architecture, including granularity, density of neurons in superficial layers, type and density of interneurons, and lateral connectivity among pyramidal neurons (Zald 2007). These marked differences are hard to reconcile with the notion that the areas carry out the same fundamental computations.

Finally, other views have, to greater or lesser extents, rejected the notion of functional specialization within PFC and instead posit that information appears distributed because function is distributed (e.g., Sleezer et al. 2016). The most extreme version of this argument, in which there is no functional organization, is not commensurate with the extensive neuropsychology literature. A more nuanced suggestion is that there are discrete, localizable processes that each contribute to a broader, integrated function of PFC as a whole (e.g., Wilson et al. 2010). This offers more parsimony with the neuropsychology literature, by accounting for differential effects of lesions, but leaves the prevailing problem that the information encoded by single units varies so little across prefrontal areas, making it hard to discern the unique components of function that occur in one area versus another.

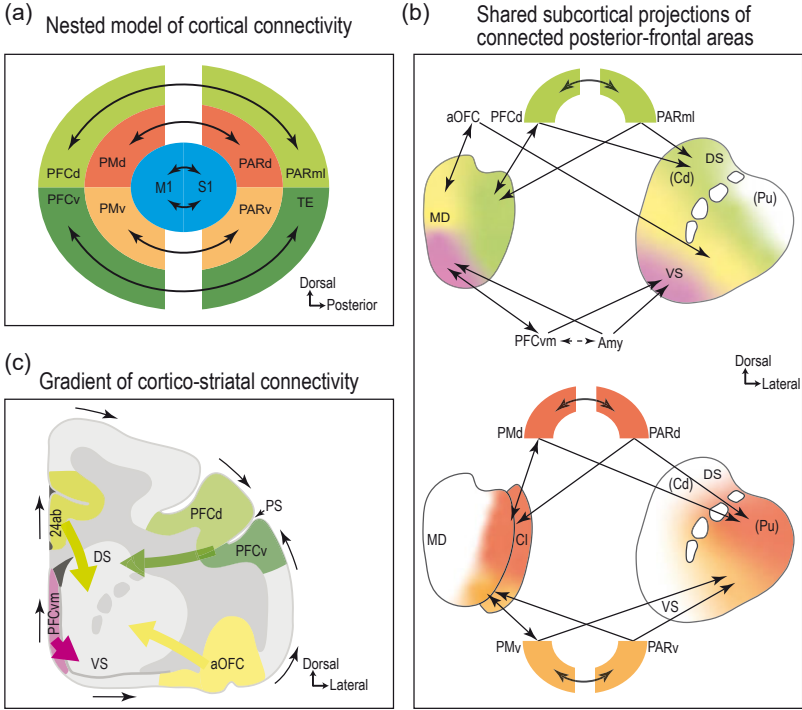
Rather than conclude that neurophysiology is homogenous across PFC or question the notion of specialization altogether, we propose two directions for reconciling the clear distinctions in neuropsychology with the relative homogeneity of neurophysiology. First, we consider the anatomical organization of PFC as it relates to larger brain circuits, where gradients of connectivity and cellular architecture are more prominent than discrete subregions. This suggests that physiological properties may also vary in a graded fashion, producing a source of variability that muddies the waters when trying to understand localization of function from the perspective of discrete regions. Second, our

standard analytic approaches that investigate neural coding might be missing the forest for the trees, and new perspectives on information coding and dynamics in neural populations could help us understand how prefrontal regions are specialized for particular functions.

### **Anatomical Organization of Prefrontal Circuitry**

The anatomical organization of PFC is an important guide to understanding its functional organization. A long history of anatomical work has fractionated PFC into discrete regions, each given a corresponding number or acronym (Brodmann 1909; Walker 1940). Early studies relied on the size and location of cells, and stains for myelin. More recent studies have used stains for increasingly complex sets of markers (Carmichael and Price 1994) or, when using imaging in humans, measures of functional connectivity (Van Essen and Glasser 2018). For the most part these studies assume that discrete regions exist, and then proceed to determine how many regions there should be and where the boundaries should be placed. Placing a boundary is always an inference process. For instance, modern techniques that use clustering algorithms use a free parameter to determine the number of clusters. Although the assumption that there are in fact discrete areas in PFC has been questioned multiple times (Kaas 1987; Lashley and Clark 1946; von Bonin and Bailey 1947), the dominant view is that discrete areas exist. Furthermore, it is assumed that each area subserves a unique function. Brodmann went so far as to assume that each area was a separate organ of the mind, with its own function. This notion of discrete areas is also reflected in the placement of lesions in neuropsychology studies.

Despite this tendency to parcellate anatomy, the balance of evidence seems to support a different interpretation. Consideration of both anatomical connectivity and comparative anatomy across species suggests that PFC can be better understood from the perspective of gradients than discrete areas with sharp boundaries. While there are some cases of clear distinctions between cortical regions (e.g., between primary and secondary sensory areas), this does not apply as well to association circuits, including PFC. The large-scale organization of PFC and related circuits has instead led to a model (Figure 5.2a) which suggests that, at the cortical level, parietal-frontal and temporal-frontal circuits are organized as nested loops, similar to an onion (Giarrocco and Averbeck 2023). At the core is primary somatosensory and motor cortex (M1/S1). At the next level there is a dorsal parietal to dorsal premotor circuit, and a ventral parietal to ventral premotor circuit. Beyond this there is a dorsal-medial parietal to dorsal prefrontal circuit, and a temporal to ventral prefrontal circuit. Although considerable anatomical complexity is not captured by this simplified model, it does capture the strongest trends in connectivity. In particular, the model articulates both hierarchical organization and specific connectivity that is likely to influence the organization of function. Furthermore, the



**Figure 5.2** Gradient organization of cortical-striatal circuits (Giarrocco and Averbeck 2023). (a) Nested model of cortical-cortical connectivity. Connectivity in neocortex is organized in a nested architecture, with posterior and frontal areas connected in circuits that show ventral-dorsal, posterior-anterior structure. (b) Prefrontal cortex connections to the striatum are organized in a gradient, such that ventral-medial and caudal orbital areas are connected to the ventral striatum, dorsolateral areas are connected to the dorsal striatum, and intermediate areas are connected to intermediate portions of the striatum. (c) Connected posterior and anterior cortical areas, in the nested architecture, have overlapping projections into the striatum and thalamus. Additionally, circuits through the striatum to the pallidum also project to similar overlapping areas in the thalamus, forming closed loops. Anterior cingulate cortex (24ab), anterior orbitofrontal cortex (aOFC), mediodorsal (MD) thalamus, dorsal parietal (PARd), mediolateral parietal (PARml), ventral parietal (PARv), dorsal prefrontal (PFCd), ventromedial prefrontal (PFCvm), ventral prefrontal cortex (PFCv) dorsal premotor (PMd), ventral premotor (PMv), principal sulcus (PS), temporal cortex (TE), dorsal striatum (DS), ventral striatum (VS).

dominant white-matter tracts linking posterior (i.e., behind the central sulcus) and anterior cortical areas connect posterior and anterior areas at the same level of the hierarchy (Yeterian et al. 2012).

When the cortical-subcortical circuitry is examined, it can be shown that the posterior and anterior nodes of these nested circuits share subcortical projections in both the striatum and thalamus (Figure 5.2c). Thus, nodes in the dorsal parietal to dorsal premotor circuit project to overlapping regions in the dorsal putamen, the lateral mediodorsal nucleus, and the adjacent central-lateral

nucleus of the thalamus. Nodes of the ventral parietal and ventral premotor circuit project to a corresponding ventral region in the same basal ganglia and thalamic nuclei. Similar overlapping projection targets can be shown for each of the connected posterior and anterior areas (Giarrocco and Averbeck 2021, 2023). Although not all connected cortical areas have overlapping subcortical projections (Selemon and Goldman-Rakic 1988), it is the case that connected areas that correspond to the nested architecture have overlapping subcortical targets.

Within this nested organization, the striatal projection target of prefrontal areas can be predicted using only the coronal and anterior-posterior locations of tracer injections (Averbeck et al. 2014), consistent with the idea that connectivity between PFC and the striatum follows a gradient. Here, the ventral-medial PFC and the caudal OFC project into the ventral striatum, the dlPFC (area 46) projects into the dorsal striatum, and areas between these two poles project to intermediate locations in the striatum on a ventral-medial to dorsal-lateral axis (Figure 5.2b). This is true whether one translates dorsomedially from ventromedial PFC or anterolaterally from OFC, toward dlPFC. A similar topography can be seen in downstream striatal projections to the pallidum and cortical and pallidal projections to the mediodorsal thalamus (Figure 5.2c). This gradient in frontal projections, combined with the overlapping subcortical projections of posterior and anterior areas, implies an overall gradient architecture in cortical-subcortical circuits.

Beyond anatomy, this model suggests an organizing principle for the functions of PFC, as well as their corresponding neurophysiological mechanisms. Specifically, there might be gradients of function within and across traditionally defined prefrontal regions. This would be important in comparisons across prefrontal regions because it would introduce a source of variability within each population, particularly if neuron sampling is wide and sparse. Indeed, when relatively large swaths of cortex are sampled at high density, graded trends are often found. For instance, spatial receptive fields in dlPFC broaden from posterior to anterior, and selectivity for objects and colors drops in a graded fashion (Riley et al. 2017; Tang et al. 2021). In contrast, moving from posterior to anterior in OFC, value encoding tends to increase (Rich and Wallis 2017). While we still lack a mechanistic picture of how these graded responses reflect an underlying function, recognizing heterogeneity can help form hypotheses of how function is organized and maps to neurophysiology.

## **Evolutionary Origins and Ventrodorsal Trends**

Evolutionary perspectives help to integrate the concepts of anatomical and functional gradients with what is known about the main divisions of PFC. Anatomical gradients and nested organization have been identified using modern tract-tracing methods in macaques, and this is also consistent with human

resting-state studies, but this architecture likely reflects the evolutionary expansion of neocortical areas. Early comparative work in reptiles identified two dominant nodes in the pallium (the vertebrate homolog of the mammalian cortex). The medial pallium is homologous to the hippocampus, and the lateral pallium is homologous to pyriform cortex. Between these nodes there are transition areas. This early work, therefore, established a tripartite model of the pallium (Abbie 1940, 1942; Dart 1934) with medial, lateral, and intermediate (possibly dorsal) areas.

Subsequent work based on developmental gene expression has extended and provided further support for this model and suggested that the pallium, and the mammalian cortex, can be divided into four regions (Puelles et al. 2017): a medial-hippocampal region, a dorsal neocortical (neopallial) region, a lateral region that develops into the claustrum and insula, and a ventral region that develops into pyriform cortex and the cortical or pallial amygdala. Whether fish, amphibians, and nonmammalian amniotes have a dorsal pallium that is homologous to mammalian neocortex is the subject of ongoing debate (Striedter and Northcutt 2020). Recent work using gene expression data from single cells has suggested that reptiles do have a neopallium, homologous to neocortex (Tosches et al. 2018). What is clear is that the neopallial region in fish, amphibians, and nonmammalian amniotes is relatively small when compared to the massive expansion of the neocortex, particularly in primates. While there has been considerable expansion in the mammalian cortex, the slope is steepest for neocortical areas (Finlay and Darlington 1995). Thus, the dorsal pallium is relatively small in nonmammalian vertebrates, relative to the medial and ventral pallium. Particularly in primates, however, the neocortex has become much larger than the medial (hippocampal) and ventral-lateral (pyriform) areas.

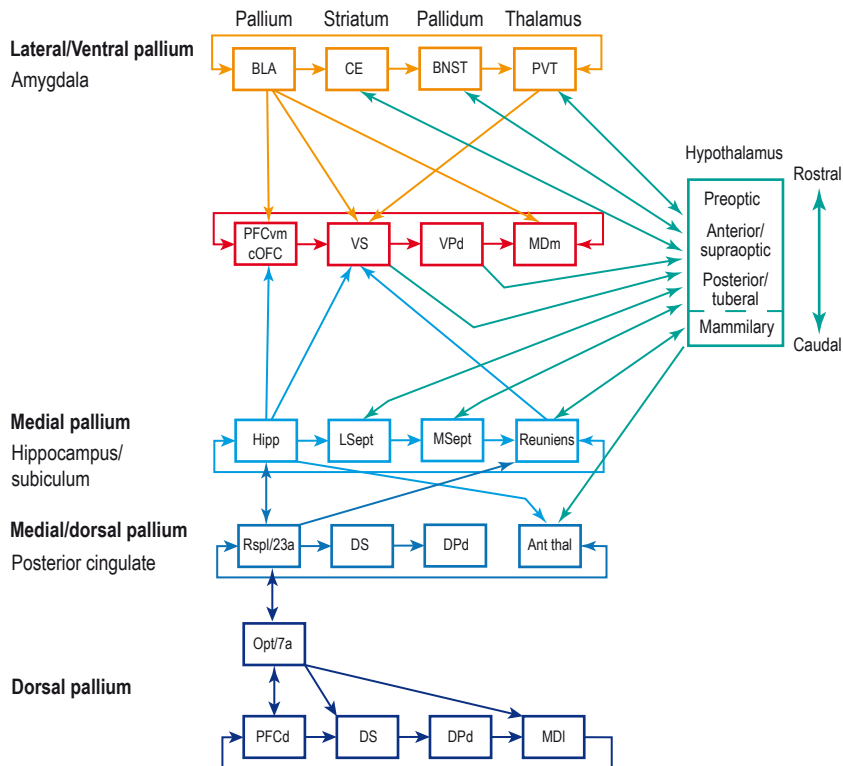
Sanides (1970) and subsequent authors suggested that the organization of PFC could be understood from the perspective of the tripartite model. In the dual-origin theory, prefrontal cortical areas expanded across evolution, as cortex expanded, starting from the medial-hippocampal and ventral-pyriform areas. The gradient anatomical organization of cortical-striatal-pallidal-thalamocortical circuits can, therefore, be understood as a topological expansion of this circuitry, from a Cambrian or possibly Precambrian ancestral vertebrate brain that was dominated by medial (hippocampal) and ventral (olfactory) circuits. As the dorsal pallium expanded, the anatomical connectivity between pallial, striatal, pallidal, and thalamic areas maintained their topological adjacency relationships as they also expanded, leading to the gradient of connectivity identifiable in primates.

The anatomical data suggests that the ancestral vertebrate brain was dominated by medial-hippocampal and ventral-pyriform pallial areas, and at most an incipient dorsal pallium. The medial and ventral pallial (cortical) allocortical areas in primates project to the ventral striatum, which projects to the ventral pallidum. The ventral pallidum is a single structure to which both the direct and indirect pathway neurons from the striatum project, similar to the pallidum

in fish and amphibians. The neocortical areas, on the other hand, project more dorsally into the striatum, which then projects to the dorsal pallidum, which is divided into internal and external segments, with direct pathway neurons in the striatum projecting to the internal segment and indirect pathway neurons projecting to the external segment. The division between internal and external segments in the pallidum, in the circuitry connected to the neocortex, is prominent in the primate.

We have previously defined the areas connected to the ventral striatum as the ventral circuit and the areas connected to the dorsal striatum as the dorsal circuit. The ventral circuitry is dominated by conserved (i.e., present across all vertebrates) medial and ventral-lateral pallial circuits, whereas the dorsal circuitry is dominated by the recently expanded neopallial circuits. The medial pallial circuits correspond to the hippocampus and the ventral-lateral pallial circuits correspond to piriform cortex. At all levels, including the cortex, ventral circuitry, similar to the classically defined limbic system, has strong connections with the hypothalamus, whereas the dorsal circuitry has minimal connections with the hypothalamus (Figure 5.3) and instead projects, via the substantia nigra, to the mid-brain tectum (i.e., the colliculus). Because the hypothalamus plays an important role in physiological homeostasis, this suggests a model where the ventral circuitry is important for identifying internal needs, and matching these needs to objects in the environment that can satisfy these needs (Averbeck and Murray 2020). The dorsal circuitry, on the other hand, is situated to use egocentric spatial information to direct actions toward objects in the environment. The ventral circuitry, therefore, establishes goals and the dorsal circuitry implements actions to achieve those goals.

This organization aligns well with lesion data and shows clear distinctions between ventral regions such as the OFC (which is important for evaluation and emotion, processes that relate to internal needs) and dorsal regions like the dlPFC (which is implicated in cognitive control used to direct attention and action). It is also echoed in the tendency of neurons in dorsal areas to encode spatial or directional information. In addition, there is some indirect neurophysiology support for separation of large-scale dorsal and ventral circuitry. Specifically, during reinforcement learning tasks in which monkeys have to learn which objects are more frequently rewarded when they are chosen, ventral circuit areas (including the cortical amygdala, orbital frontal cortex, and the ventral striatum) maintain a representation of the values and identities of behavioral goals between trials and during baseline hold periods before choice options are presented (Tang et al. 2022a). Presumably, this value- and goal-related information (in the form of the representation of the to-be-chosen visual stimulus) reflects a match between mechanisms in the hypothalamus that code thirst or hunger depending on the unconditioned reinforcer used in the experiments and the visual stimulus on the screen. Further, it has been found that, when the choice options are presented, the value and identity information flows into dorsal circuits where it is used to identify and direct an action toward



**Figure 5.3** Organization of pallial-striatal-pallidal-thalamo-pallial circuits (Giarrocco and Averbeck 2023). Lateral and medial pallial areas are strongly connected to the hypothalamus, whereas recently evolved dorsal pallial areas have minimal connectivity with the hypothalamus. BLA: basal-lateral amygdala, CE: central nucleus of the amygdala, BNST: basal nucleus of the stria terminalis, PVT: paraventricular nucleus, PFCvm: ventromedial prefrontal cortex, cOFC: caudal orbital frontal cortex, VS: ventral striatum, VP: ventral pallidum, MDm: medial portion of the medial dorsal thalamus, Hipp: hippocampus, LSept: lateral septum, MSept: medial septum, Rspl: retrosplenial, DS: dorsal striatum, DPd: dorsal pallidum, Ant Thal: anterior thalamic nuclei, PFCd: dorso-lateral prefrontal cortex, MDI: lateral portion of the medial dorsal thalamus.

the spatial location of the object (Tang et al. 2022a). The value and identity information is not, however, strongly represented in the dorsal circuit during the intertrial interval or other periods in the task when actions cannot be planned or directed to goal objects. This hypothesis was motivated by a consideration of the anatomical circuitry, and specifically by differential connectivity between forebrain circuits and the hypothalamus. The current neurophysiological data that supports the hypothesis is based on a stronger representation of the behavioral goal, which in this case is a visual stimulus, during the intertrial interval and initial fixation period in the ventral circuitry, and a stronger representation of the actions, at the time of choice, in the dorsal circuitry (O’Reilly 2010).

## **Population Coding and Dynamics**

In addition to anatomical organization, functional dissociations may be obscured by the methods used to analyze neural responses. Historically, prefrontal neurophysiology has focused primarily on the activity of single neurons and identifying the experimental factors that change their firing rates. More recently, as it has become common to record many neurons simultaneously, efforts have increased to understand how information is represented at the population level and how computation is performed over such representations. Although information can be extracted from single neuron firing, these neurons are embedded in interconnected networks, both local and long range. Therefore, it may be more accurate to conceptualize neuron responses that we record as snapshots of activity in a larger dynamical system. If population perspectives have increased validity over single unit analyses, they may also be able to reconcile the disconnect between neuropsychology and neurophysiology in the search for functional specialization.

Similar to single units, encoding properties can be assessed in neural populations. One might analyze how activity varies across time or conditions, when the functional unit is not a single neuron but a population of neurons. This can be done by considering each neuron as an axis in a high-dimensional space. For instance, if we record the activity of 100 neurons, the population response can be considered as a 100-dimensional representation that evolves over time, with any time window characterized by a vector of 100 firing rates. By doing this, the response of any given neuron is necessarily considered in relation to others in the population, so that information is not represented in the activity of any one neuron, but as a pattern of activity over the population. From this starting point, one can take multiple approaches. If the population is sampled repeatedly under different conditions, classifiers can extract task information from the population vectors by differentially weighing elements (neurons). Similarly, population dynamics can be captured by the path the vector takes through the high-dimensional neural state space. Repeated samples of these paths define the region of neural space in which activity resides, referred to as a manifold.

Because neural activity is not random and includes shared variance, the population activity that defines a manifold usually exhibits structure and is lower in dimensionality than the theoretical potential of a sampled population (Gao et al. 2017). That is, a good deal of variance in our 100-neuron population might be captured by only a few dimensions. Dimensionality reduction finds dimensions of shared variance, allowing us to understand whether they correspond to task or cognitive variables. Heading direction, for example, is a two-dimensional variable. Thus, activity in circuits representing heading direction might reside on a two-dimensional manifold, perhaps nonlinear, in population coding space. Given multiple samples of a population under different conditions, shared variance across samples could be found agnostically with



an approach like principal component analysis. Projecting the original samples onto the first principal component summarizes the original data in a single dimension, or “subspace,” and allows us to ask whether activity in that subspace varies across conditions. In this case, subspace is a generic term referring to a lower-dimensional linear projection of population activity, defined by applying weights to each neuron in a population vector. These weights might be determined in a number of ways. While a principal component analysis captures the axes of maximum variance in population, they may be poorly aligned with the dimensions in which task conditions vary. Therefore, an alternative subspace might be defined by axes oriented to condition-wise variance. In any of these reduced-dimensionality spaces, one can assess how dynamics evolve and vary with experimentally defined conditions. Indeed, much of this decrease in dimensionality may have to do with the relative simplicity of tasks used to study neural activity (Gao et al. 2017). For example, if a population response is a (potentially nonlinear) mapping from task variables into population coding space, then low-dimensional tasks will necessarily lead to low-dimensional population activity. By extension, more naturalistic tasks that include many dimensions of variability are expected to increase the dimensionality of neural representations. However, the dimensionalities of populations in natural conditions are not yet clear, in part due to the challenges of interpreting behavior and neurophysiology in unconstrained tasks.

Because population approaches afford a different perspective on neural coding, they may provide unique insights into how representations, and the computations over these representations, vary across PFC. For instance, although task-relevant information tends to be encoded by single neurons throughout PFC, different features may be emphasized by different populations, such as expected rewards in OFC and cognitive variables in dlPFC. An example of this is population activity that creates a geometry where the relevant condition on a trial is clearly distinguished by a large separation of different conditions in state space, with other task-relevant information embedded in that structure (Chien et al. 2023). Large separations can lead to a form of abstraction, in which different instances that share a common feature occupy nearby or overlapping regions of the neural state space, which may allow the concept to generalize to new instances (Bernardi et al. 2020). Such possibilities can be investigated by evaluating the geometry of population representations.

In another domain, population dynamics traverse different landscapes, the features of which could vary in different PFC regions. For instance, dynamics in PFC often tend toward consistent dynamical trajectories, fixed points, or other attractor basins. These are believed to be stable points in the neural activity space that may be formed by patterns of synaptic weights within a network (Averbeck 2022). Therefore, attractor states could be influenced by both intrinsic architecture and experience-related plasticity, both of which could vary across PFC regions. Importantly, these dynamics arise from the collective activity of a group of neurons, so that any one unit might reflect some

fragmented features but is unable to reveal the overall picture. For instance, in lateral prefrontal (prearcuate) cortex, population dynamics separate sensory inputs from the computation of an upcoming choice, even though these are intermixed at the single neuron level (Mante et al. 2013). In OFC, population dynamics reveal transient representations of two choice options that alternate during deliberation, where single units only revealed the chosen option (Rich and Wallis 2016). Although these studies each focused on one region at a time, cross-regional comparisons that use similar techniques could help us better understand the neural mechanisms that support unique functions within and across regions of PFC.

### **Summary and Open Questions**

It is widely believed that there is functional specialization within PFC, so it is natural to expect neurophysiology to provide clarity on the nature of the unique function of a region. To date, however, this clarity has not emerged. Instead, single neurons tend to represent “everything everywhere.” Although these data demonstrate the flexibility of prefrontal neurons, they have so far failed to reveal major differences between neurons recorded from different regions. In light of this, we have highlighted two considerations for future studies. First, evolutionary and anatomical data suggest two dominant trends within PFC, each with gradient-like organization that is more prominent than discrete boundaries. Investigating neural coding with respect to this anatomy may be fruitful for understanding local and global organization of function. Second, examining the representations and temporal dynamics that emerge from neural populations may provide unique insights into local function and cross-regional interactions. One recent study has taken steps in both of these directions, by using population representations to assess the flow of information across lateral PFC. Here, information flowed in the caudorostral direction when the location of a valuable object needed to be identified, and in the dorsoventral direction when preparing an eye movement to that location (Tang et al. 2021). Approaches such as these hold promise in revealing how populations represent and communicate information.

In addition to the approaches highlighted here, there are others that could provide important insights. In particular, a defining feature of different prefrontal regions is their unique patterns of connectivity, and approaches aimed at understanding interactions among interconnected regions could reveal key differences. One way to accomplish this is to combine perturbation studies with neural recording. A study that did this found that neurons in both OFC and ACC encode reward values, but only OFC neurons showed altered value coding following amygdala lesion (Rudebeck et al. 2013a). Similarly, studies that quantify functional connectivity between regions can determine how PFC interacts with targets elsewhere. To the extent that these interactions differ

between PFC subareas, these approaches may also shed light on functional specialization.

Although we have suggested avenues for future investigation, there are still many open questions. Population approaches are increasingly popular in neurophysiology, yet it remains to be determined whether they will ultimately provide unique insights into functional specialization. To this end, we need to know which anatomical or functional properties define a population. This may be particularly challenging to address if functions are graded, meaning discrete boundaries do not apply. Populations in nonhuman primate recordings are often samples of opportunity, defined by the access the researcher achieved and limited by current recording methods. However, if high-density recordings are collected along the entire anterior-posterior length of the principal sulcus, should they be analyzed as one population or many, and if the latter, where should divisions be drawn? The rapid advance of technology, in terms of the scale and type of recordings we can collect, presents new opportunities to address these questions. In addition, population approaches are usually agnostic to neuron type, connectivity, or laminar location, none of which are typically known when neurons are recorded from nonhuman primates. Yet methods for identifying subtypes of neurons or their projections or recording across cortical laminae are becoming more prevalent, which means that we should soon be able to evaluate some of these questions rigorously. Taken together, although the neurophysiological distinctions among prefrontal regions are not obvious and neural encoding appears superficially similar, there is reason to be optimistic that pursuing in-depth understanding of anatomical organization and neural coding may help parse the neurophysiological mechanisms that distinguish the fundamental functions of different regions of prefrontal cortex.

