12

Integrative Psychological, Computational, and Mechanistic Approaches to Frontal Lobe Function

Amitai Shenhav, Marie T. Banich, Christian Beste,
Timothy J. Buschman, Naomi P. Friedman,
Caterina Gratton, Etienne Koechlin, Nicolas Schuck,
Xiao-Jing Wang, and John O'Doherty

Abstract

Since the earliest accounts of the prefrontal cortex (PFC), its core functions have remained elusive and hotly debated. Here, an attempt is made to bring order to these varied accounts and to account for the heterogeneous observations that have been made across methodologies and species. After cataloging the myriad functions that have been attributed to PFC and the approaches that have been taken to taxonomize these functions, a new framework is proposed for conceptualizing PFC function. This framework is based on a set of four canonical computations that is argued to collectively provide a more formal, coherent, and comprehensive account of existing findings regarding PFC function. These canonical computations include goal-directed integration, active maintenance, selection of task-relevant information, and monitoring. Discussion includes how previous PFC findings can be understood through one or more of these functions, and ways in which these computations may collectively form a motif that repeats throughout regions of PFC over different forms of inputs and outputs. Finally, critical directions for future research to validate or falsify this account of PFC functions are highlighted, including the leveraging of new and emerging directions for experimentation and analysis.

Group photos (top left to bottom right) Amitai Shenhav, John O'Doherty, Marie Banich, Naomi Friedman, Timothy Buschman, Nicolas Schuck, Etienne Koechlin, Caterina Gratton, Christian Beste, Xiao-Jing Wang, Marie Banich, Amitai Shenhav, John O'Doherty, Timothy Buschman, Christian Beste, Etienne Koechlin, Caterina Gratton, John O'Doherty, Xiao-Jing Wang, Nicolas Schuck, and Naomi Friedman

From "The Frontal Cortex: Organization, Networks, and Function," edited by Marie T. Banich,
Suzanne N. Haber, and Trevor W. Robbins. 2024. Strüngmann Forum Reports, vol. 34,
Julia R. Lupp, series editor. Cambridge, MA: MIT Press. ISBN 9780262549530

What Adaptive Functions Does Prefrontal Cortex Serve?

A Starting Point to Encapsulate PFC Function

Any integrative account of a given brain structure is destined to be incomplete and in need of revision, particularly when that brain structure subsumes the entirety of the prefrontal cortex (PFC). Even if the puzzle is not likely to be fully solved, one can at least strive to start to integrate as many of the critical pieces as possible. The challenge is knowing whether one is starting in the right place to solve the puzzle with the pieces that one has in hand, or whether one needs to start from scratch.

In seeking to compile and bring order to the functional mechanisms underpinned by the PFC, we will, therefore, start by asking: What is the broad range of psychological functions and phenomena in which this structure has been implicated? We then proceed to consider different approaches to taxonomizing and/or decomposing this array of functions, and what these taxonomies collectively reveal about possible canonical computations that unify or at least reduce the dimensionality of PFC functions. Finally, we discuss how the study of PFC function and its underlying computations can be improved by extending traditional methods and leveraging emerging experimental, analytic, and modeling approaches.

There are several sources of data that researchers have taken into account when attributing functions to the PFC (Table 12.1), including

- Cognitive impairments observed in individuals with PFC damage (e.g., lesions), inactivation (e.g., cooling or other noninvasive brain stimulation methods), and/or deterioration (e.g., frontotemporal dementia) of the PFC,
- PFC functions that are altered over the course of evolutionary development (across species) and/or ontological development (particularly over early development) along with development and maturation of these structures, and
- PFC functions whose engagement covaries with increased neural activity within and/or across prefrontal regions (as measured, e.g., via electrophysiology or neuroimaging).

Functions Commonly Ascribed to the PFC

Active Maintenance

Working memory is the ability to actively maintain a limited set of information in the absence of direct sensory input for short periods of time (e.g., 3–10 seconds). It is critical for complex cognition, allowing one to break free from the immediate world (i.e., simple stimulus-response associations) and to keep critical information at the ready. Working memory has been considered

From “The Frontal Cortex: Organization, Networks, and Function,” edited by Marie T. Banich,
Suzanne N. Haber, and Trevor W. Robbins. 2024. Strüngmann Forum Reports, vol. 34,
Julia R. Lupp, series editor. Cambridge, MA: MIT Press. ISBN 9780262549530

Table 12.1 Semi-exhaustive list of functions commonly attributed to PFC.

Cluster	Sample Functions
Active maintenance	• Maintaining goals, values, task-relevant cognitive and emotional information • Buffering goals from interference
Selection	• Selecting/determining goals based on internal and external context • Distinguishing relevant vs. irrelevant information in the environment • Selecting relevant information from memory • Selecting specific information for prioritization • Emotion regulation and reframing
Versatility	• Suppressing prepotent responses (e.g., habits) • Shifting between goals or tasks • Arbitrating between hypotheses and strategies • Flexibility to novel, unfamiliar, or changed environments
Monitoring	• The environment for task-relevant information • Whether the correct action has been selected • Whether one's action led to the desired goal • Whether goals and actions align with values
High-level combinatorial processing	• Abstraction, generalization • Identifying novel or atypical strategies/solutions • Coordinating goals, learning, and memory • Constructing value • Processing for multiple tasks • Language • Reasoning
Simulation	• Envisioning novel solutions or courses of action • Simulating forward or backward in time • Hypothesis testing • Metacognitive processing • Social inference (e.g., theory of mind)

one of the canonical functions of the PFC. Decades of research, starting with groundbreaking recordings from Fuster and Goldman-Rakic, have promulgated the idea that the contents of working memory are actively maintained or referenced in the pattern of neural activity within the PFC, most notably in the face of distraction. Working memory is both capacity and time limited, enabling the maintenance of about 4–7 items for time periods up to about 10–15 seconds. Despite these limitations, it is highly flexible with regard to content. One can hold any type of information (e.g., verbal, spatial, emotional) in working memory, and neural correlates are likewise flexible in what they can represent. Neurons (or neural populations) in PFC have been found to actively represent sensory inputs (Fuster and Alexander 1971; Romo et al. 1999), motor actions (Mars and Grol 2007), the value or emotional significance of stimuli (Platt and Padoa-Schioppa 2008; Rolls et al. 2009; Salzman and Fusi 2010), actions (Barraclough et al. 2004; Shin et al. 2021), and task rules (Wallis et al. 2001; White and Wise 1999).

Selection

PFC has been implicated in selecting those representations and processes that are most relevant for the current task goals (e.g., Miller and Cohen 2001). For example, PFC may bias toward processing of specific relevant attributes of the external world (e.g., color, portions of space; Banich et al. 2000; Kastner and Ungerleider 2001) or types of information (e.g., linguistic; Snyder et al. 2014), memory (e.g., semantic; Wang et al. 2018), actions (e.g., action sequences; Zhang et al. 2021), emotion regulation (e.g., reappraisal; Braunstein et al. 2017), or abstract plans (e.g., steps required to traverse a subway system; Balaguer et al. 2016), all of which are selected in reference to current task goals.

The putative role of PFC in selection has also been exemplified in impairments observed during the selection of options in decision-making tasks. For instance, classic lesion studies in humans implicated ventral PFC (including the orbitofrontal cortex, OFC) in the selection of stimuli associated with varying reward values, especially following changes or reversals in reward associations (Murray et al., this volume; Bechara et al. 1997; Fellows and Farah 2003; Hornak et al. 2004; Noonan et al. 2010). There is accumulating evidence to suggest that ventral prefrontal regions, especially the OFC, may be especially important for selecting between stimuli based on the prospective rewards associated with them, whereas more dorsal parts of the PFC, including dorsal anterior cingulate cortex (ACC) and pre-supplementary motor area, may play more of a crucial role in making decisions over actions (Aquino et al. 2023; Camille et al. 2011b; Rudebeck et al. 2008b). PFC also appears to play a role in selecting between more abstract policies (e.g., different strategies, expert systems), which we discuss further below.

Versatility of Responding and Thought

Here we consider two aspects of the versatility of responding and thought: overcoming habitual patterns of responding and being able to switch flexibly between responses or thoughts. In terms of the former, let us consider Teuber's description of behaviors associated with frontal lobe damage, which he characterized as "bewildering" in variety (Teuber 1972:637) yet sharing elements of "compulsiveness" or "abnormally stimulus-bound behavior" (p. 640). That is, individuals with frontal lesions might be unable to avoid habitual responding in a given context in favor of less automatic responses which might be more appropriate in that context. Moreover, stimuli in the environment can trigger automatic responses; for instance, seeing a computer will engender starting to type on the screen (Lhermitte 1986; Lhermitte et al. 1986). This suggests that one ability enabled by the frontal lobes may be the ability to respond to stimuli in different ways beyond the stereotypical manner based on well-learned responses.

Another aspect of flexibility is the ability to change one's course of action or thought processes. Such a switch may be driven by external information that signals a change in certain processes is now possible or desirable, or with regard to external feedback about the utility of those processes under the current context or an internal evaluation of the efficacy of actions. Such abilities are compromised in individuals with damage to the frontal lobe (e.g., Adólfsdóttir et al. 2014; De Baene et al. 2019).

Monitoring

Critical to ensuring that one's actions and choices are efficacious in leading to a goal, one must evaluate or monitor outcomes or internal states, as in emotion regulation or memory retrieval. Monitoring refers to how an agent tracks its own behavior and/or the consequences of those behaviors in various situations (i.e., in the face of information obtained from the environment), which can impose varying demands on behavioral control (Botvinick et al. 2004; Holroyd and Coles 2002; Rushworth et al. 2004). Such monitoring processes depend on medial and superior PFC activity (Giller et al. 2020; Reinhart and Woodman 2014). Activity in these regions is increased in situations that are unexpected or deviate from one's goal (e.g., error commission). This suggests that increased monitoring during such situations is required to enable behavioral control. Such monitoring abilities are compromised after frontal lobe damage (e.g., Hochman et al. 2015). Importantly, the degree of monitoring has to be balanced to be able to cope with changes in situational requirements. This dynamic balancing in the degree of cognitive control monitoring has been termed "meta-control" (Eppinger et al. 2021; Hommel and Wiers 2017) and shown to be altered by disorders affecting frontal lobe functions, such as in obsessive-compulsive disorder and attention-deficit hyperactivity disorder (Colzato et al. 2022).

Higher-Level Combinatorial Processing

There is evidence that information maintained and selected by PFC can reflect a higher-level combination, or abstraction, of current sensory input or internal representations. Studies have, for instance, shown that categorization, which often requires a nonlinear combination of sensory variables, involves the PFC (e.g., Freedman et al. 2001; Seger and Miller 2010). Other studies have shown that populations of PFC neurons are engaged when animals switch between tasks that require the animal to focus on different aspects of the same stimulus

(Mante et al. 2013). These neurons show two important coding characteristics: (a) they exhibit mixed selectivity, meaning that a cell can be responsive to multiple cognitive features (e.g., Aoi et al. 2020); (b) they appear to be able to reduce information to underlying dimensions, such as being able to code information in discrete categories (e.g., Mack et al. 2020). The process of abstraction has also played a central role in research on value-based decision making (e.g., Cortese et al. 2021; De Martino and Cortese 2023), with orbitofrontal regions of PFC being implicated in representing "partially observable" information, such as context from past events, in the service of maximizing reward (Schuck et al. 2018; Wilson et al. 2014).

Simulation

Planning—a function known to be impaired after damage to PFC (Owen et al. 1990; Shallice and Burgess 1991)—relies on simulation. Simulation describes a process of bringing to mind (or "sampling from") potential future states of one's environment, including the potential positive or negative consequences of arriving in this state. The mental representation of these potential future states and outcomes is referred to as a world model. By mentally sampling a world model, one can identify valuable and efficient courses of action. As Sutton and Barto (1998) point out, this form of learning (model-based reinforcement learning, RL) can be equivalently viewed as moving forward into potential future states or as revising backward the courses of action which led to such states. It has thus been thought that PFC is critical for managing/controlling covert simulated behavior in the same way as overt behavior (Campbell et al. 2018). Additional evidence for this role, which we elaborate on later in our discussion of unifying features, comes from findings that regions throughout PFC track information related to the value of current and future states, as well as how these values are transformed to guide behavior.

Summary

We recognize that this listing of functions is likely not exhaustive. It also does not identify any new processes that have not been discussed previously in the literature. Nonetheless, it does identify core functions that involve the full extent of frontal regions.

Existing Approaches to Divide the Space of PFC Function

What Do We Want a Taxonomy of PFC Function to Accomplish?

The groupings offered above provide one form of functional taxonomy, but one whose boundaries are defined arbitrarily. To develop a better taxonomy, it is important to ask first what sorts of properties are needed to make such a taxonomy useful and effective. In other words, what are the criteria by which one might determine that they succeeded or failed in developing a good taxonomy of PFC function?

The first property that one might seek in a taxonomy of PFC function is its *descriptive* utility: How well does it capture variability in PFC function within an individual over time, and across individuals? To what extent does it capture deficits reported by PFC-damaged patients? How does it align with variability in prefrontal anatomy and physiology, including patterns of functional activation and connectivity? To what extent does it capture variability in PFC-related behavior, function, and structure over the course of development or in response to stressors?

The second property that one might seek is its *generative* utility. Can it be described in formal terms, and at a level of description that can be assessed across species and methods? Does it give rise to new assays (e.g., new tasks, metrics) that allow researchers to capture more precisely the sources of variability above? Does it identify ways of applying existing measures (e.g., behavior, physiology) and interventions (e.g., inactivation, pharmacology) to those assays to test new hypotheses? Does it point toward targeted treatments that alleviate deficits in patients with damage or dysfunction linked to PFC?

Forms of Taxonomy: Strengths and Limitations

Qualitative Description of Behavioral Impairment

Taxonomies drawn from observations of behavioral impairment after frontal lobe damage have a long and storied history, starting most famously with the case of Phineas Gage, a railroad construction foreman whose crew was excavating rock in 1848 to build a railroad line in Vermont. While using a tamping iron to pack an explosive into a borehole, a spark from the iron on the rock detonated the explosive, leading the rod to pierce the anterior portion of his left frontal lobe through the eye socket (Macmillan and Lena 2010). In the oft quoted description, changes in both social and cognitive characteristics were noted afterward. Socially he was no longer sensitive to others and could be profane, and while previously he had held the position of a construction foreman, he could no longer come up with a plan and systematically follow through on it. Other individuals who have suffered from frontal lobe damage in modern times have exhibited deficits on self-reports of their ability to deploy executive functions successfully in their daily lives (e.g., Løvstad et al. 2012).

Task Impairment

A more quantitative and systematic approach to understanding PFC-related impairments has focused on mapping out those regions where damage through lesions is commonly implicated in aberrant performance on well-characterized laboratory tasks (e.g., Godefroy et al. 2023; Meier et al. 2022). Such studies have a number of strengths and limitations. With regard to strengths, any taxonomy of frontal lobe function from patients is arguably most relevant for real-world behavior, as alterations to frontal lobe function are observed across a wide variety of neurological and psychiatric disorders. On the other hand, there is potential for reorganization of function between the time of damage and assessment. Moreover, lesions often span important morphological and functional boundaries in the brain, which can make determinations difficult and/or preclude studies from having large numbers of participants with damage to one particular brain region.

Factor Analysis of Performance across Tasks

As discussed by Duncan and Friedman (this volume), factor analysis has been used to evaluate whether performance on executive function and so-called "frontal lobe" tasks are influenced by a single or multiple underlying factors of ability. This question emerged from models of working memory, which suggested a central executive that controlled the contents of storage buffers (Baddeley 1986). In seeming contradiction to the notion of a unitary executive, executive function tasks showed low correlations. However, low correlations could arise even if there were a unitary central executive because executive tasks show low reliability and "task impurity." Because executive functions control other processes, executive tasks must include these non-executive functions, as differences in these can also influence performance (Miyake et al. 2000). Thus, Miyake et al. (2000) selected sets of tasks intended to tap three executive functions—response inhibition, working memory updating, and mental set shifting—that varied in these lower-level processes and used confirmatory factor analysis to extract latent variables. Latent variables are based only on shared variance across a set of tasks, so they can remove random measurement error as well as variance due to non-executive demands that differ across tasks (i.e., task impurity). They found that these latent variables showed moderate correlations, suggesting some shared variance, or "unity," but these correlations were significantly lower than 1, suggesting some distinct variance, or "diversity," even after accounting for task reliability issues. Thus, their conclusions, which were based on a sample of neurally intact college students, echoed conclusions of earlier studies that focused on frontal lobe damage (Duncan et al. 1997; Teuber 1972), which suggested "unity and diversity" of frontal lobe function.

Although this study might be described as creating a "taxonomy," it is important to note that Miyake et al. (2000) never intended this battery to capture "core" or "elemental" components of executive functions. They decided to focus on these three functions because they were among the most commonly examined executive functions at an intermediate level of analysis, but they explicitly noted that other executive functions likely existed and that functions could be conceptualized at different levels (e.g., planning might be composed of multiple sub processes). This study illustrates the principle that taxonomies can exist at multiple levels depending on the researchers' goals; this set of functions provided a tractable means with which to tackle the goal of evaluating whether commonly hypothesized executive functions could be considered unitary. That said, this has also proved useful in subsequent research to evaluate the relations of unity and diversity components to other constructs of interest, such as other cognitive processes, psychopathology, and neural areas (see Friedman and Miyake 2017).

A parallel approach is to utilize meta-analytic tools to find terms that are commonly associated with activation in prefrontal regions. For example, using a topic modeling approach, de la Vega et al. (2016, 2018) found that certain terms (e.g., inhibition, conflict, working memory, and decision making) are associated with studies that yield prefrontal activation. Terms could then be examined to determine with which regions of medial (de la Vega et al. 2016) and lateral (de la Vega et al. 2018) frontal cortex they are associated.

Theory-Driven Decomposition of Function

The set of functions attributed to PFC can be decomposed into interlocking functions that can be described along one axis by their control "effectors," that is, the distinct sets of controlled processes that are subsumed by each. For instance, different forms of control can be described as involving selective enhancement of particular processing streams (e.g., forms of selective attention; Desimone and Duncan 1995), directed search, and retrieval of information held in episodic or semantic memory (e.g., cued recall, prospection; Polyn et al. 2009; Schacter et al. 2008), transformation of information held in working memory (e.g., mental rotation, inference; Olivers et al. 2011; Shepard and Metzler 1971), and parameterizing one's decision process (e.g., response threshold; Bogacz et al. 2006; Leng et al. 2021; Wiecki and Frank 2013). Each of these define different forms or *types* of control that one can engage, many of which have been linked to regions of PFC (Duncan 2010; Miller and Cohen 2001; Shenhav et al. 2013, 2016).

However, the presence of these controllers alone is incomplete without an account of when, why, and *to what degree* (i.e., with what level of *intensity*) each of these are selectively engaged, disengaged, or modified (Hommel and Wiers 2017). Thus, an orthogonal level of functional description needs to provide at least a minimal account of the process by which each type of control is (a) selected (i.e., determining the appropriate amount/s and type/s of control to allocate), (b) executed (i.e., engaging the relevant control processes), and (c) monitored (i.e., identifying conditions under which control needs to be adjusted) (Botvinick and Cohen 2014).

Unlike the factor analytic approach described above, this form of functional taxonomy does not derive directly from quantitative task behavior, which is a limitation of the approach. It does, however, serve a similar purpose in providing a coherent lower-dimensional structure to the set of processes underpinning performance across those tasks. These taxonomies instead derive or take inspiration from a combination of psychological, neural, and/or computational evidence and are refined by the same (e.g., evidence of the neural and computational distinctiveness of different forms of control, dynamics of post-error performance adjustments). This approach serves as both a strength (drawing from convergent sources of evidence) and a limitation (affords a high level of subjectivity and flexibility in how to weigh the strength and plausibility of different sources of evidence).

Neurobiological Fractionation

A contrasting domain of approaches focuses on subdividing processes based on neurobiological criteria. These may be functional (e.g., fMRI or electrophysiological activations in particular tasks, correlations of functional signals across regions, changes in functional responses after damage to particular brain areas) or anatomical (e.g., macro-anatomic based on sulcal morphology or connectivity of major tracts, or micro-anatomic based on cytoarchitecture, receptor densities). Some common themes have emerged from this work, including the presence of specialized brain regions, and evidence that these brain regions join together to form large-scale brain networks (e.g., see chapters by Vertes et al., Gratton et al., and Murray et al. in this volume).

An advantage of these approaches is that they can provide a new way of conceptualizing

- Divisions in prefrontal function and constraints on theories of function (e.g., regarding the unity and diversity of functions or the types of processes that can be plausibly represented by neurobiology),
- How these divisions arise (e.g., via ties to evolution, development, and plasticity of neurobiology), and
- How different forms of brain damage can be biologically represented (e.g., via ties to particular regional functions, neurotransmitter modes of actions, models of interregional connectivity).

For example, as reviewed by Gratton et al. (this volume), resting-state functional connectivity has been shown to subdivide the cortex, including the PFC. In these descriptions, 10–17 networks are identified with fairly distinct spatial organization. The frontoparietal, cingulo-opercular/salience, default mode (A and B), dorsal attention, and ventral attention/language are the most studied "association" systems of PFC (see discussion in Gratton et al. on taxonomy and visualizations of these networks). The clear modularity exhibited by these networks (with high within-network connectivity and low between-network connectivity, replicable across groups, people, over time within a person) suggests that these may underpin fractionable functions within the frontal cortex. Indeed, these networks are associated with dissociations in task responses, anatomical features, electrophysiological response properties, neurodegenerative disease, and predictive of behavioral performance. These distinctions become even clearer in individual-level mapping that addresses issues of inconsistent spatial localization across people.

However, a limitation of these approaches is that they are largely descriptive and not closely tied to a mechanistic understanding of PFC function or cognition. While component processes identified with behavioral/cognitive measures show some overlap with neurobiological subdivisions (e.g., Duncan and Friedman, this volume), their alignment is not always clear; see discussion of cingulo-opercular and frontoparietal function in Gratton et al. (this volume). Thus, while neurobiological findings help to constrain theories, they may provide limited insights on their own regarding how functions are implemented in the PFC and give rise to differences in behavioral outcomes.

Computational Models: A Tool for Formalizing Taxonomies

One challenge for cognitive taxonomies is that verbal descriptions of functions are often vague, which can make them less useful for making predictions. Computational models can address this issue by recapitulating core aspects of behavior while providing a more formal, more reproducible, and less ambiguous description of the different functions, hence enabling quantitative predictions about behavioral and/or neural changes that result from arbitrary manipulations. Another challenge for most taxonomies is that a focus on the behavioral versus cognitive versus biological level can yield different results, while leaving unclear the translation between the different taxonomies. Computational descriptions could allow us to identify links between the different levels and help provide a mechanistic understanding that bridges the biological and cognitive levels, as illustrated by a recurrent circuit model of working memory and decision making (Wang 2002).

Computational models integrate multiple operations into a consistent functional system that can be used to investigate the empirical performance of individuals performing tasks described by that system. These simulations can then be used to test whether a model can reproduce subjects' behavior along with related neural activity, and to compare the degree to which distinct models can reproduce such empirical data so as to identify key computational operations within a consistent integrated system.

One of the advantages of a computational approach is that it can provide a common language that helps us bridge multiple levels of understanding and measurement. Computational models can, for instance, make predictions at the network level, about activation of a broad region, about patterns of neural activity within a region, and/or about distributions of receptors. In this way,

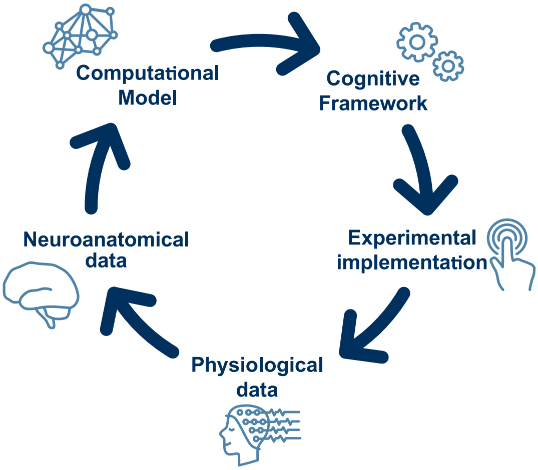

Figure 12.1 Interdependencies between theoretical and experimental approaches to investigating PFC function. Computational models help shape and formalize conceptual and theoretical frameworks for understanding cognition. Together, these serve to operationalize and form testable hypotheses, inspiring specific experiments for measuring relevant neural function and structure. Data collected from such experiments, in turn, serve to constrain preexisting models and/or adjudicate between multiple alternative models.

having a strong computational framework can allow researchers working with different methods and at different levels of description to communicate and inform one another.

Importantly, computational models work in tandem with and are directly informed by other levels of investigation (see Figure 12.1). A clear conceptual (e.g., cognitive) framework is necessary to allow the field to connect the insights gained from a computational model to the conceptual background that has been around in the field for a long time and has inspired well-validated experimental procedures. Physiological data (e.g., electrophysiological recordings during the experiment) and information about the neuroanatomy can then be further used to inform the computational approach taken.

How Can Psychological, Neurobiological, and Computational Approaches Constrain One Another?

Naturally, the process of identifying candidate functions, constructing computational models of those functions, and then mapping those functions onto biology is iterative and multidirectional. Identifying neurobiological mechanisms of prefrontal function will likely improve our understanding of what functions are important for cognition and how these are implemented in computational models of cognition. Conversely, identifying and specifying cognitive functions associated with executive control can motivate the design

From "The Frontal Cortex: Organization, Networks, and Function," edited by Marie T. Banich,
Suzanne N. Haber, and Trevor W. Robbins. 2024. Strüngmann Forum Reports, vol. 34,
Julia R. Lupp, series editor. Cambridge, MA: MIT Press. ISBN 9780262549530

of new computational models (e.g., flexible working memory) which can, in turn, generate testable hypotheses for how these functions/computations are accomplished in the brain.

A good computational model of the frontal lobe must be able, for instance, to account for the presence of the known dissociable functional network architecture present in this region of the brain. Such a modular organization may emerge as a product of the modeling approach or may be necessary to implement to constrain the model. Different networks may act on different forms of information (indexed by their connectivity) but use similar canonical computations (discussed below). Alternatively, canonical computations may differ across networks (e.g., perhaps between the language and frontoparietal/multiple demand system).

Functional and neurobiological methods can also provide an understanding of the types of factors that a good model must be able to account for as well as estimates of their range/variability. For example, even in the normative population, substantial interindividual variability has been observed in executive function performance (Duncan and Friedman, this volume), in the spatial layout and extent of functional brain networks (Gratton et al., this volume), and in sulcal morphological characteristics (Murray et al., this volume).

What Are the Unifying Features of PFC Function?

What Are Essential or Canonical Computations Within PFC?

Next, we turn to understanding the canonical computations underlying the adaptive functions of the frontal lobe. Integrating theories drawn across the many taxonomies described above, we identified a set of four putative canonical computations performed by PFC (see Figure 12.2):

1. *Goal-directed integration* involves the ability to access, combine, and sequence information so it can be used effectively to create goals and subgoals, and is supported by the diverse anatomical connections of the frontal lobe, which allow it to integrate information across all cognitive domains.
2. *Maintenance of information* involves the ability to actively maintain representations over time, which supports the ability of the brain to sustain goals and direct cognition.
3. *Selection of task-relevant information* allows for the selection of information and representations, especially at a more abstract level, that are most relevant for current goals.
4. *Monitoring* enables the ability to compare expectations to outcomes, including the prediction of future outcomes, which enables the ability to monitor cognition and flexibly adapt to a changing world.

From “The Frontal Cortex: Organization, Networks, and Function,” edited by Marie T. Banich,
Suzanne N. Haber, and Trevor W. Robbins. 2024. Strüngmann Forum Reports, vol. 34,
Julia R. Lupp, series editor. Cambridge, MA: MIT Press. ISBN 9780262549530

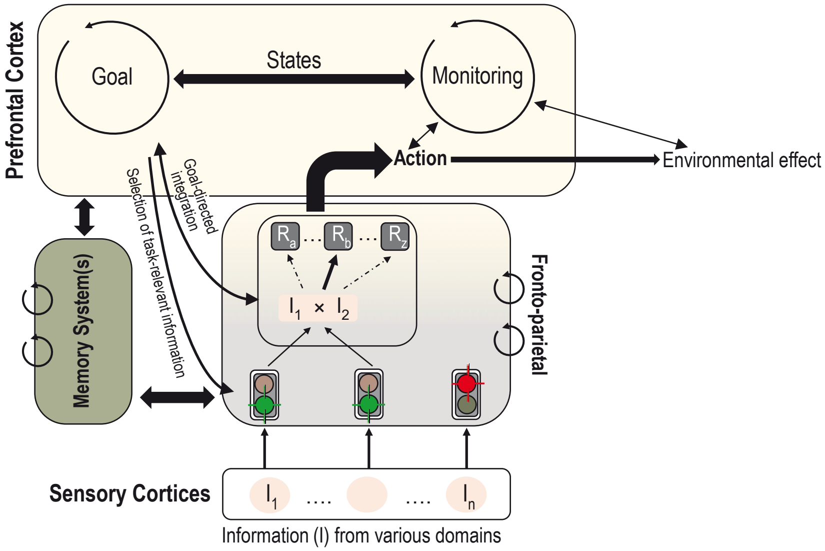

Figure 12.2 Illustration of canonical computations applied to an example of a cognitive task. Information about sensory features of current stimuli must be integrated to determine the appropriate response based on one's current goals and task set. This process requires actively maintaining representations of relevant stimulus features, actions, and/or goals in working memory. Goal-driven processes may act to bias processing of certain features or responses, particularly in cases where automatic processing of those features promotes responses inconsistent with one's current goal. Information indicating deviations from one's goal (e.g., errors, processing conflict) is monitored to modify ongoing and future control (e.g., biasing).

Before we briefly describe each of these canonical computations, it is important to note that these core computations were selected to be parsimonious: a simple set of functions that encompass the broad range of adaptive functions listed above. These functions can be applied broadly to give rise to cognition, across a variety of inputs, cognitive domains, and timescales. As we detail below, these computations do not act alone; complex cognition can only arise through the dynamic interaction and sequencing of these computations.

Goal-Directed Integration

A major innovation over the last few decades of research on PFC function was the proposal of a multiple demand (MD) system (Duncan 2010): a common set of brain regions in frontal and parietal cortices that are active across a variety of different cognitively demanding tasks. The MD system consists of distinct patches that can be found in both hemispheres and which span the lateral prefrontal regions, insular cortex, the dorsomedial frontal cortex, lateral

From "The Frontal Cortex: Organization, Networks, and Function," edited by Marie T. Banich,
Suzanne N. Haber, and Trevor W. Robbins. 2024. Strüngmann Forum Reports, vol. 34,
Julia R. Lupp, series editor. Cambridge, MA: MIT Press. ISBN 9780262549530

and medial parietal cortices as well as temporal regions (Assem et al. 2022). As discussed by Duncan and Friedman (this volume), "with parts widely distributed through the cortex, strongly interconnected with one another, the core MD system is well placed to take in and integrate representations of many kinds and flexibly feed out the results for selective cognitive control," a process dubbed "attentional integration" (Duncan et al. 2020). Here, we expand on this conceptualization to identify specific forms of integration that occur within PFC.

Integration of sensory and motor representations. The PFC receives parallel streams of sensory and motor information. Superior parietal mechanisms contribute to the selection of motor responses (Bernier et al. 2012; Cisek and Kalaska 2002; Jaffard et al. 2008), possibly because the superior parietal cortex plays a central role in stimulus-response translation processes (Gottlieb 2007). There is, however, a well-known "binding problem" of how the sensory representations become connected to motor representations. To resolve this problem, the theory of event coding (TEC) (Hommel 2004; Hommel et al. 2001) draws on common coding principles to put forward the concept of an event file, which reflects the integrated representation of sensory and motor features that are themselves stored in distinguishable representations. According to TEC—and more recent derivatives thereof, which also consider functional neuroanatomical structures and neurophysiological mechanisms (Beste et al. 2023)—the coding and dynamic handling of event files involves structures in the parietal and PFC that strongly overlap with brain regions that constitute the MD system (Duncan 2010). Thus, commonalities between different instances of executive functions may become explainable through a smaller set of (computational) mechanistic principles relating to the integration of sensory and motor task sets.

Numerous lines of evidence suggest that the coding of integrated sensory and motor representations involves inferior and superior parietal areas, supplementary motor areas, the dorsolateral PFC, and the hippocampus (Chmielewski and Beste 2019; Dilcher et al. 2021; Kleimaker et al. 2020). Superior and posterior parietal areas integrate perception and action by binding sensory information into a common representation of the association between stimuli and responses (Gottlieb 2007). In a similar vein, regions of the temporoparietal junction contribute to this process by using environmental information to update these mental representations (Geng and Vossel 2013). So, through parietal mechanisms, the PFC is presented with different options for how to respond. The PFC then likely has to decide which of the different options to use and to connect with the appropriate motor program or task set that leads to observable behavior.

Integration of goals, values, schemas, memories, affect, and actions/policies. A primary challenge for the brain is to integrate the numerous aspects that make

From "The Frontal Cortex: Organization, Networks, and Function," edited by Marie T. Banich, Suzanne N. Haber, and Trevor W. Robbins. 2024. Strüngmann Forum Reports, vol. 34, Julia R. Lupp, series editor. Cambridge, MA: MIT Press. ISBN 9780262549530

up cognitive functioning, such as goals, strategies, values, affective states, actions (and their affordances), sensory inputs and observations, and existing memories. We suggest that one canonical computation of the PFC lies in integrating these complex levels in a manner that serves to produce goal-oriented behavior or thought. This process entails integrating higher-level variables, such as one's goals and current affective state, to produce a course of action that best achieves these goals, which in turn can lead to changes in internal states (e.g., selective memory retrieval) and execution of particular action plans. This integration function is closely related to the ability to form complex combinations, as discussed above, and in particular to the idea that complex decision-making tasks require abstractions that can be thought of as a cognitive map or task set.

Task sets describe the relevant sensory information, representations, and actions needed to meet a specific goal under specific conditions. By analogy with the hippocampus, which has been shown to integrate multiple cortical representations into episodes (Eichenbaum 2017), a task set can be viewed as a large-scale neural frame integrating multiple representations distributed over cortical regions (e.g., stimulus-action mappings, action-outcome predictive models) that can be evoked collectively to form a consistent system that guides behavior. These task sets can, in turn, enable the PFC to regulate adaptive behavior. This notion of task sets or rules in PFC also relates to theories of RL discussed earlier, wherein it is proposed the PFC encodes a rich set of world models (e.g., of how objects and agents in our environment might interact). These world models can be flexibly applied to new situations via a probabilistic inference process about their relevance (Tomov et al. 2023; Tsividis et al. 2021).

The task sets that result from this integration process are closely linked to value signals and outcomes of RL in the brain (e.g., Schuck et al. 2016; Wilson et al. 2014), which also have been widely observed in ventromedial PFC (Adelhöfer and Beste 2020; Beierholm et al. 2011; Hampton et al. 2006; Hardung et al. 2017; Lee et al. 2014; Narayanan et al. 2013), but arguably extend to striatal and other brain areas (Sharpe et al. 2020). Some work has suggested that the computational function unique to the PFC, in particular the OFC, is to provide representations that go beyond merely observable information by adding relevant information of the past (context) (Niv 2019; Schuck et al. 2018; Wikenheiser and Schoenbaum 2016). It should be noted, however, that the integration performed by PFC goes beyond these processes and includes, for instance, integration of information across different strategies (e.g., Donoso et al. 2014b) and expert systems (Charpentier et al. 2020; Lee et al. 2014; O'Doherty et al. 2021). Moreover, the temporal scale across which integration is performed can be much longer than a single task, allowing the emergence of meta-learning.

Robust/Active Maintenance of Information across Functions

Maintaining information is critical to a wide array of cognitive functions. Classic studies focused on the maintenance of sensory inputs or the preparation of motor actions (Funahashi et al. 1989, 1993a; Fuster and Alexander 1971). This maintenance allows cognition to break free from the immediate world, integrating information over time and responding at the appropriate time. Active maintenance of information, however, is also critical for more "cognitive" variables, such as maintaining information about the current situation, the current task, one's goals, and the value of different options.

To support the integration functions above, different types of information must be integrated over many different timescales; while a current thought is only briefly maintained, goals can extend longer, from a few minutes of focusing on writing a manuscript to years of dutifully saving for retirement. These different timescales of integration are reflected in the variety of intrinsic timescales of individual neurons. The variety of timescales found in the frontal lobe may reflect the diversity of functionality; neurons with shorter time constants respond to stimulus inputs while neurons with longer time constants maintain that information in working memory (Wasmuht et al. 2018).

It is important to note that the maintenance of information is not passive. Rather, it is focused on task-relevant information. Part of the reason for this feature is that working memory has a severely limited capacity: we are able to hold only a few items (i.e., 4–7) "in mind" at once. Therefore, it is important for selection mechanisms to determine what information is allowed to enter working memory, often referred to as "gating" (O'Reilly and Frank 2006; Yang et al. 2016), as well as mechanisms to select individual memories to drive behavior, akin to attention to external stimuli (Gazzaley and Nobre 2012; Panichello and Buschman 2021). Of note, it has been shown that during such "gating" processes, similar brain regions and neurophysiological processes are in charge that are also relevant for the integration of sensory and motor representations, but via different pathways of information processing that terminates in frontopolar regions (Yu et al. 2022). Beyond overcoming limitations in capacity, focusing the contents of working memory on task-relevant information can also ensure that only goal-relevant information is represented, integrated, and acted upon, and that extraneous information does not intrude or interfere. This function requires a further type of canonical computation: selection.

Selection and/or Biasing/Regulation of Task-Relevant Information

The world is incredibly rich. At each moment in time, we are inundated with a flood of sensory information from the outside world: potential memories we could recall, thoughts we could manipulate, actions we could take. Filtering this flood is critical to cognition. It allows us to focus our behavior on those stimuli/memories/actions that are contextually relevant. Filtering also helps to focus learning on those representations that are believed to be important, helping to resolve which features of the environment are most predictive of potential outcomes (referred to as the credit assignment problem). These considerations suggest that a "selection" mechanism is a canonical computational function of PFC.

Selection Over Representations of the Outside World. Attention is perhaps the best studied form of selection. Decades of research suggests PFC plays a central role in internally directed attention. Neurons in PFC represent where attention is allocated in space and to what features (Buschman and Kastner 2015; Miller and Cohen 2001). Activity in these prefrontal regions is observed prior to activity in other brain regions, suggesting PFC plays a leading role in directing attention (Buschman and Miller 2007). Stimulating within PFC induces attention-like effects in visual cortex (Moore and Armstrong 2003), and inhibiting/lesioning causes deficits in tasks requiring attention (Bichot et al. 2019).

Attention acts to filter cognition by biasing representations in other brain regions. For example, directing attention to a spatial location increases the activity of visual cortex neurons with receptive fields at the attended location (Reynolds et al. 2000). This increase in activity acts through lateral inhibition to suppress other competing representations (Desimone and Duncan 1995; Reynolds et al. 1999; Reynolds and Heeger 2009). In this way, attention can selectively focus sensory processing on a subset of neural representations. Several alternative mechanisms have been proposed to achieve the same effect: synchronizing the activity of neurons can increase their impact on downstream neurons (Fries et al. 2001), decreasing noise correlations can improve the signal-to-noise ratio (Cohen and Maunsell 2009), and changing the geometry of neural representations may allow certain information to flow between brain regions (Panichello and Buschman 2021). In the end, top-down guided selection acts likely through a confluence of mechanisms to filter information in other brain regions.

Selection over internal representations. Selection is not limited to attention to sensory inputs. It can also act in other domains. For example, frontal cortex plays an important role in controlling recall from episodic memory. As reviewed by Eichenbaum (2017), animals and humans with prefrontal damage have trouble selectively recalling information from episodic memory because of intrusion of competing memories. This suggests that although PFC does not provide direct mono-synaptic inputs into the hippocampus, it plays an important role in selective recall from episodic memory. This also refers to the selection of integrated sensory-motor representation, which are also thought to be stored in episodic traces (Hommel 2009).

Selection can also filter representations within frontal cortex. As noted above, selection is critically important for protecting the limited capacity of working memory. A "gating" mechanism is thought to control what information enters working memory (O'Reilly and Frank 2006; Yang et al. 2016). Then, when multiple items are held in working memory, an "internal attention" mechanism acts to select one item and use it to guide behavior. Functional imaging has shown PFC regions that direct attention to internal representations (in working memory) also direct attention to external, sensory representations (Gazzaley and Nobre 2012). Consistent with these findings, recent electrophysiological recordings in monkeys show the same neural representation encodes both selection from working memory and sensory inputs (Panichello and Buschman 2021). Similar to selective attention, selecting an item from working memory biases the neural representation to improve the encoding of the selected item. These findings suggest that control representations in PFC may be domain-general, allowing the brain to select task-relevant information regardless of the source of information.

Selection of higher-order cognitive variables, including goals and information processing parameters (meta-control). Finally, selection may also act on higher-order cognitive representations that can influence neural processes themselves. For example, research has shown that people will adapt the parameters of learning and decision making depending on the current context (e.g., changing the decision threshold or affecting the time constant of integration) (Cavanagh et al. 2011; Dayan 2012; Leng et al. 2021; McGuire et al. 2014). These forms of "meta-control" may occur through the biasing of competition between potential strategies (O'Doherty et al. 2021) or by direct selection/adjustment of parameters governing the relevant learning and decision processes. Electrophysiological recordings suggest that this form of control adjustment may happen by selection acting on different cortical regions, for instance, amplification of neural representation in order to filter representations appears to occur in sensory cortex, while adjustment to decision criteria have been localized to the frontal cortex and/or basal ganglia (Beste et al. 2018; Cavanagh et al. 2011; Forstmann et al. 2008; Frank et al. 2015; Luo and Maunsell 2015, 2018).

The role of inhibition in selection. Inhibition is inherent in the concept of selection. Selecting one item is, by necessity, to the detriment of other representations. Projections from the frontal lobe are largely excitatory (although see interhemispheric inhibition in mice; Cho et al. 2023). This suggests inhibition occurs through local mechanisms in the circuit that is receiving the selection signals. One such mechanism would be local lateral inhibition (e.g., through parvalbumin-positive inhibitory interneurons; Cardin et al. 2009). In this way, selection can act positively to strengthen selected representation which would, in turn, act through lateral inhibition to suppress other representations. In the field of attention this mechanism is often referred to as the "biased competition model" (Reynolds and Heeger 2009), although it can be generalized to other domains (Carandini and Heeger 2012). It has also been argued that "inhibition" may be the byproduct of the top-down biasing done by the PFC, such as by maintaining a task set or goal, because such biasing does so to the detriment of other representations (Munakata et al. 2011).

Alternatively, selection may act through direct feedforward inhibition that specifically suppresses a particular representation or region. Such mechanisms may be important for inhibiting responses, thoughts, or memory recall (Depue et al. 2016; Hulbert et al. 2016).

With either mechanism, varying the strength of the inhibition could modulate the strength of selection. Moderate selection could allow multiple representations to co-exist, but with a bias toward the selected representation(s). In contrast, strong inhibition could lead to winner-take-all dynamics that select a single representation, which may be important when only one response can be emitted (Wilson et al. 2012).

Monitoring

Timescales. Monitoring is a key dimension of control. Monitoring processes evaluate the relevance and reliability of behavioral policies and cognitive strategies guiding behavior to identify the need to inhibit, enhance, or revise them to make behavior more adaptive and efficient. Monitoring processes are likely distributed over the PFC and have been proposed to operate on three main temporal dimensions: (a) *retrospectively* from actual action outcomes to reactively adjust control processes guiding ongoing behavior (e.g., within vmPFC or dACC), (b) *prospectively* from contextual cues to proactively adjust control processes before acting (e.g., within lateral PFC), and (c) *counterfactually*, regarding alternative behavioral policies/strategies that are not guiding ongoing behavior but might advantageously replace the current behavioral policy/strategy guiding ongoing behavior (e.g., within frontopolar PFC, Koechlin and Wang, this volume).

Sources. Dorsomedial PFC, including the dorsal ACC and the pre-supplementary motor cortex has long been found to encode error or conflict signals during performance of complex tasks. These signals were first observed in EEG studies in which the so-called error-related negativity has been found, localized to dorsomedial PFC, which has been argued to be related to an internal detection that an error has occurred (Fu et al. 2023; Gehring et al. 1993; Hauser et al. 2014). Similar error signals have also been found to occur at the time of feedback. One possible source of these error signals is the reward prediction error (Holroyd and Coles 2002; Schultz et al. 1997), which detects discrepancies between expected and actual outcomes, possibly reflecting the effect of dopaminergic innervation into medial frontal cortex. These kinds of error signals have also been found to be present in both pre-SMA and anterior cingulate neurons in both monkey and human studies, as well as in BOLD fMRI in humans (Debener et al. 2005; Phillips and Everling 2014; Shen et al. 2014; Wang et al. 2005).

From an electrophysiological perspective, all these signals share a reliance on theta oscillations in the medial frontal cortex, which, due to biophysical principles, are optimally suited to integrate information being processed in distant brain regions (Buzsáki and Draguhn 2004; Cavanagh and Frank 2014). These theta-related processes are believed to reflect a "surprise signal," indicating a need to adapt one's actions (Cavanagh and Frank 2014), for instance, when the executed action mismatched the correct one. These kinds of signals are thought to be important for providing a metric of how well one is performing on a task, whether it is in terms of successfully getting rewards or implementing intended actions. For this evaluation to occur, it is relevant to rely on a comparison process, according to which information about the expected effects or the action plan need to be retrieved—a process likely guided by theta as well as gamma band information (Beste et al. 2023). It is possible that these processes reflect integrated representations of stimulus and action features (Beste et al. 2023), and that these integrated representations reflect content-specific beta band activity, which changes from active to latent to reactivated states as needed (Spitzer and Haegens 2017; Wendiggensen et al. 2022). The interplay of theta and beta related activity is likely under the control of alpha band activity to flexibly balance between top-down and bottom-information (Beste et al. 2023; Wendiggensen et al. 2023). It is the interplay of these oscillatory activity patterns that is likely central for above-discussed canonical computations, referring to perceptual and motor task sets (Beste et al. 2023), and which may also give rise to dynamics and functions reflected within PFC and the broader MD system (Duncan 2010).

These monitoring signals are also likely important for facilitating changes in strategy. Reliability is another form of signal that is important for monitoring and evaluation, which goes beyond the punctate-based error signals based on single events. Reliability concerns how well a particular strategy is doing in terms of making predictions and can be considered to be related to the (inverse of) variance or degree of uncertainty in the predictions associated with a particular strategy (Daw et al. 2005; Lee et al. 2014; O'Doherty et al. 2021). One way to compute reliability is by integrating over prediction errors; for instance, if many reward prediction errors have occurred recently, then reliability of reward predictions can be said to be low, whereas if only few small errors have occurred, we can say that reward prediction reliability should be high. Reliability signals for different strategies (such as for model-free vs. model-based RL strategies or even between different ways of learning through observation) have been found to correlate with BOLD responses in ventrolateral PFC and frontopolar cortex in humans (Charpentier et al. 2020; Lee et al. 2014; O'Doherty et al. 2021), whereas reliability signals related to different possible model-based strategies (i.e., within the model-based system) have been found in ventromedial PFC and frontopolar cortex (see Koechlin and Wang, this volume). Thus, PFC appears to monitor performance at different levels of abstraction, from punctate error signals to strategy reliability signals.

Targets of adjustment. Another way in which forms of monitoring dissociate relates to the type of control they are supporting. Research has shown that a hierarchical gradient of control emerges in the lateral PFC, with more caudal areas of lateral PFC representing information lower in this hierarchy and more rostral regions representing information higher up in the hierarchy (Badre and D'Esposito 2009; Badre and Nee 2018; Koechlin et al. 2003). It was subsequently proposed that parallel regions along the medial wall may engage in forms of monitoring that subserve control at similarly increasing levels of response complexity (Taren et al. 2011; Venkatraman and Huettel 2012). For instance, caudal regions of dorsomedial PFC (potentially corresponding to the cingulo-opercular network) have been shown to track the amount of conflict between competing responses (e.g., should I respond left or right), whereas more rostral regions of dorsomedial PFC (potentially corresponding to the frontoparietal network) have been shown to track the amount of conflict between potential strategies or other higher-order goals (e.g., should I maintain my current strategy or switch) (Ritz and Shenhav 2024; Shenhav et al. 2018; Venkatraman et al. 2009a).

How Do These Canonical Computations Align with Behavior and Neurobiology?

Alignment with Behavior

Any one task likely involves all of the canonical computations outlined above: integration, maintenance, selection, and monitoring/evaluation. However, each task may place a different distribution of demands on these computations. As a result, the relative contribution of PFC to each of the relevant computations might vary across tasks. There can, for example, be tasks in which the monitoring/evaluation aspect takes more prefrontal computational resources than the other canonical computations or where this is the case for integration, maintenance, or selection. For instance, a typical response interference-based cognitive control task (e.g., Stroop, flanker, go/no-go) may place limited demands on integration of task-relevant information (e.g., linking stimulus features with appropriate responses) and/or maintenance (e.g., of relevant task rule), but greater demands on monitoring (e.g., for errors or processing conflict) and/or selection (e.g., biasing of task-relevant feature processing). Conversely, for a typical decision-making task (e.g., choosing between foods, goods, or gambles), the demands on goal-directed integration may be more substantial,

From "The Frontal Cortex: Organization, Networks, and Function," edited by Marie T. Banich,
Suzanne N. Haber, and Trevor W. Robbins. 2024. Strüngmann Forum Reports, vol. 34,
Julia R. Lupp, series editor. Cambridge, MA: MIT Press. ISBN 9780262549530

requiring comparison across values of relevant features of the options and potential courses of action (Frömer and Shenhav 2022).

This involvement of multiple canonical computations with differences in their relative weighting might lead to observations of unity and diversity of functions (e.g., in individual differences in performances across tasks).

Alignment with Neurobiology

Differences in function versus differences in representation. It is possible that the various functions above are subsumed by distinct regions of PFC. Alternatively, it is possible that there are canonical computations that are repeated across subregions within PFC but with differing inputs and outputs. For example, there may be a cortical or subcortical circuit motif that actively maintains a representation. As noted above, this mechanism is broadly useful for sustaining stimulus, motor, or task representations. Therefore, the same circuit motif operating on differing inputs could serve different functions. This might explain observations of functional differences between regions (see Murray et al., this volume). For instance, spatial information is represented more strongly in lateral PFC than OFC, which may reflect anatomical differences in connectivity with parietal inputs to lateral PFC and insular, temporal, and amygdalar inputs to OFC (see Rich and Averbeck, this volume). Similarly, gradients in abstraction along the rostral-caudal axis may reflect positioning along the cortical hierarchy (Badre and D'Esposito 2007; Badre, this volume). Computational modeling has shown that repeating circuit motifs in a hierarchical structure, such that the output of one circuit feeds into the next, can describe the increase in time constants observed along the cortical hierarchy (Murray et al. 2014; Koechlin and Wang, this volume).

One advantage of this theory is that it is easier to conceptualize how the functional diversity within PFC could evolve or develop. Rather than needing mechanisms to generate unique circuits for different functions, the same circuit motif could be "copy-pasted" but still support different cognitive functions.

Which anatomical distinctions are less well-aligned with these computations? There is currently some debate as to whether specific regions of PFC are *not* specialized for the domain-general processes described above, but rather for more domain-specific processing, more specifically language. For 150 years, portions of the left inferior frontal cortex have been associated with language output. While some theories posit that the left inferior frontal gyrus is important for domain-general processing of relational and sequencing information (Fitch and Martins 2014; Pallier et al. 2011), others have argued that the left inferior frontal gyrus is organized such that these domain-general regions are interdigitated with more language-specific regions (Fedorenko and Blank 2020).

From "The Frontal Cortex: Organization, Networks, and Function," edited by Marie T. Banich,
Suzanne N. Haber, and Trevor W. Robbins. 2024. Strüngmann Forum Reports, vol. 34,
Julia R. Lupp, series editor. Cambridge, MA: MIT Press. ISBN 9780262549530

Another outstanding question is how the above taxonomy of frontal lobe function explains the functions of those portions of the frontal lobe that are associated with the default mode network. While the functions of some of these default mode regions are accounted for by the functions described above (e.g., value calculation by portions of ventromedial prefrontal function), exactly what the function of, for example, lateral DMN regions (e.g., area 8) is, and how they may or may not fit into the above taxonomy, remains unclear.

Computational Modeling in Interplay with Experimentation

Computational Building Blocks and Cross-Level Understanding

Computational modeling has always been an integral part of research on PFC function (Cohen et al. 1996) and has traditionally often distinguished between so-called algorithmic and implementational levels of modeling (cf. Marr 1982). We propose that the time is ripe to eschew this distinction and to conceptualize instead PFC-related models in terms of computational building blocks, their biological mechanisms and computational principles as laid out in the previous section. For some of these core processes, such as internal maintenance of working memory or time integration in decision making, it is possible to achieve cross-level understanding from cell types to recurrent neural population dynamics to behavioral performance (Arnsten et al. 2010; Goldman-Rakic 1995). For other, more complex cognitive functions, the underlying biological mechanisms remain poorly understood. Nevertheless, at a minimum, modeling serves as a tool, in close reciprocal interaction with experiments, to bridge phenomenological description at one level and explanation at another level.

Internal Maintenance and Manipulation of Information

Neural circuit models based on neurobiology have been developed for working memory and decision making (Wang 2002), suggesting a "cognitive-type" local circuit model of the PFC (Wang 2013). A neural network model can be designed by intuition or shaped by training using machine-learning algorithms. In the latter case, how the function is realized is not defined *a priori*; it emerges as a result of training connection weights, for instance, using a backpropagation algorithm. Building such a model for working memory-dependent tasks revealed that self-sustained persistent activity is necessary when information must not only be maintained but also manipulated to perform a task (Masse et al. 2019).

Such a model was designed to enable mechanistic understanding across multiple levels, with collective neural population dynamics described as attractor states providing an account of function/behavior, on the one hand, and enabling investigation of underlying cellular and molecular mechanisms on the other hand. In particular, a gating mechanism for filtering out distractors was proposed in terms of a microcircuit motif composed of three types of inhibitory neurons (Wang et al. 2004b). The dependence on NMDA receptors for recurrent excitation (Wang 1999) provided one clue as to why NMDA receptor signaling pathology might cause cognitive deficits in schizophrenia, one of the findings that prompted the emergence of computational psychiatry (Redish and Gordon 2017; Stephan and Mathys 2014; Wang and Krystal 2014).

Extending recurrent neural network models to rule-based tasks, such as the Wisconsin Card Sorting Test, led to the theoretical proposal of mixed selectivity of neuronal function (Rigotti et al. 2010). This was supported by experimental data (Rigotti et al. 2013) and suggests a computational advantage of complex neural firing patterns commonly observed in the PFC (Fusi et al. 2016).

Task Set Representation

The novel approach of training recurrent neural networks (Yang and Wang 2021) has also been used to realize a single network capable of performing many rule-based cognitive tasks (Bouchacourt et al. 2020; Yang et al. 2019). This approach makes it possible to investigate how task sets are represented, the (di)similarity between neural representations of different tasks, and suggests clues as to how the PFC may represent various task sets (Sakai 2008).

Monitoring and Evaluation of Performance and Outcomes

Monitoring and evaluating one's behavior in the service of task performance or learning is a fundamental aspect of PFC function. In many of the tasks discussed above, monitoring and evaluation reflects a continuous learning process that shapes future behavior based on previous outcomes. RL models have been the primary framework to computationally understand this monitoring and learning processes. At the heart of RL models is a process that monitors how achieved outcomes compare to expected outcomes and updates future expectations accordingly. RL models have been widely studied and validated as a model of the brain and behavior. Importantly, they can go beyond a simple outcome monitoring process in multiple ways, for instance, by incorporating cognitive maps that provide the model with planning abilities or by including state-inference or state-learning processes that can map observations onto abstract representations or learn the abstractions suitable for reward maximization, as is the case in deep Q network.

Within research on cognitive control, monitoring has been instantiated as a comparator that accesses information from a neural network-like architecture (e.g., levels of coactivation across response units), and it uses the result of this comparison process to modify ongoing processes across the network (Botvinick et al. 2001; Botvinick and Cohen 2014; Holroyd and Coles 2002). Recent work has augmented these monitoring algorithms to weigh additional factors relevant to the organism, including expected reward rate within the current environment and resource limitations, such as effort costs (Musslick and Cohen 2021; Shenhav et al. 2013). Cutting across research on decision making and cognitive control, an emerging theme has been increased focus on how such monitoring processes can be leveraged toward arbitrating between high-level action plans—between, for example, different strategies (Donoso et al. 2014b), model-based versus model-free decision making (Daw et al. 2005; Lee et al. 2014), expert systems (O'Doherty et al. 2021), or gradual versus state-inference based learning (Zika et al. 2023).

Large-Scale Cortical Network Model

Graph theoretical approaches have been used to analyze and model data associated with large-scale networks of the brain. For example, graph models can be developed for large-scale brain network architecture based on either structural or functional connectome data, and these models can then be lesioned in silico to generate predictions regarding the consequences of different forms of brain damage (Alstott et al. 2009; Honey and Sporns 2008; Sporns 2016). These models can then be tested to see whether their predictions are consistent with findings from human lesion patients (Gratton et al. 2012). Recent investigations have focused on multilayer modeling to represent linked changes in brain networks over time (Betzel and Bassett 2017; Gerraty et al. 2018; Muldoon and Bassett 2016), using control system models to form predictions about how different functional states can arise from a static structural connectome (Gu et al. 2015), and using dynamic oscillator models to link transient "events" to the development of a modular network architecture (Pope et al. 2021).

Using connectomic data, dynamical models have been developed for the large-scale primate cortex, both for monkeys (Chaudhuri et al. 2015) and humans (Deco et al. 2014; Demirtas et al. 2019). Among findings from this new line of research are the concept of macroscopic gradients of biological properties (Wang 2020) and a hierarchy of time constants along the cortical hierarchy (Chaudhuri et al. 2015; Murray et al. 2014), offering a mechanistic explanation for the PFC's capability of time integration in contrast to early sensory areas, which lack such a temporal mechanism. This model can be used to computationally explore how the PFC works together with the rest of the cortex, such as in working memory (Froudist-Walsh et al. 2021; Mejias and Wang 2022; Wang 2022).

Integrating across Modeling Approaches

Mutually Constraining Models across Levels of Detail

One way in which these various modeling approaches can be better integrated is by extracting information from mechanistic models and linking it to network models. This approach might be fruitful in the domain of individual differences. One could use computational models of executive processes to estimate individual-level parameters (e.g., learning rates), after which one could examine whether such parameters are associated with characteristics of brain networks. For example, one might hypothesize that individuals with faster learning rates show greater integration of information in the frontoparietal control network from various sources, as would be reflected in a higher value for the graph theoretic measure of participation coefficient.

Conversely, one might examine neural network models for properties expected based on graph models of brain function, such as the presence of large-scale modules and connector hubs (e.g., Gratton et al., this volume). These correspondences could be used as a criteria for model selection or incorporated more explicitly into model creation.

Incorporating Observations about Finer-Grained Structure

Functional brain organization differs systematically among individuals on a number of dimensions, including brain network topography, topology, areal size, and even morphological characteristics such as tertiary sulci (Gordon and Nelson 2021; Voorhies et al. 2021). Many of these differences have been linked to differences in brain function, such as task activations (DiNicola et al. 2020; Gordon et al. 2017; Seitzman et al. 2019; Tavor et al. 2016) and are predictive, in the sense of cross-validation, of individual differences in behavioral performance (e.g., Finn et al. 2015; Kong et al. 2019). It is unclear, however, why differences in the size, shape, or location of brain regions should necessarily be linked to performance. What processes benefit from access to additional neurons or particular neural circuits? Linking these observations of individual differences in structure and morphology to neural network models, such as local circuit models (Wang 2022), may provide additional deeper insights into the links between brain network organization and behavioral outcomes.

How Can These Models Be Used to Understand Unity and Diversity?

Confirming Mapping between Task Measures and Function

One benefit of models that formalize a given set of functions is that they allow you to simulate behavior on a given task and ask to what extent different parameters map onto different sources of variability in task performance. They also allow you to invert this process and ask to what extent a given measure of task performance selectively taps into a function of interest. For instance, Musslick et al. (2019) examined to what extent various common cognitive control task measures reflected individual differences in control capacity (i.e.,

From "The Frontal Cortex: Organization, Networks, and Function," edited by Marie T. Banich,
Suzanne N. Haber, and Trevor W. Robbins. 2024. Strüngmann Forum Reports, vol. 34,
Julia R. Lupp, series editor. Cambridge, MA: MIT Press. ISBN 9780262549530

how much control a hypothetical person might be able to maximally apply within a task), something that clinicians and developmental researchers often seek to index. These authors simulated a variety of task performance metrics for a given agent, including differences in performance between trials that (a) are incongruent versus congruent (congruency effect), (b) follow an incongruent versus congruent trial (conflict adaptation), and (c) follow a change versus repetition in task rule (switch costs). By simulating task performance across an array of artificial agents varying in control capacity as well as other model parameters (e.g., learning rate, task automaticity), they showed that the congruency effect, commonly used to tap into individual differences in capacity, is more likely to reveal individual differences in automaticity than capacity. At the same time, these theoretical analyses also revealed task measures that may provide a more sensitive measure of capacity (like conflict adaptation effects) and revealed more generally the extent to which these different parameters are likely to be confused with one another when using a given task measure. This approach can be extended to any of the modeling approaches described above, to aid in selection and development of tasks targeting different computations of interest.

Understanding Frontal Lobe Function through the Lens of Artificial Intelligence

As artificially intelligent agents evolve in the direction of generalized intelligence, they will likely have to overcome many of the same computational problems faced by the biological brain. The expansion of the frontal lobe over evolution has allowed for the expansion of cognition (Weiner et al., this volume). Therefore, it seems reasonable to expect that aspects of the evolution of cognition in artificial agents will involve the expansion of the same computational mechanisms that are served by the frontal lobe. Indeed, this is already reflected in many of the advances in artificial intelligence (AI) over the past several decades. Early neural network models were built using simple individual neurons with strict feedforward connectivity. While these networks were sufficiently flexible to capture complex cognitive processes, they were notoriously difficult to train to perform complex tasks. As techniques evolved, the introduction of recurrence allowed these networks to capture temporal dynamics and, importantly, begin to maintain memories of recent inputs. The next critical insight came from the introduction of selection-like mechanisms, whether it is gating of inputs into recurrent networks, such as long short-term memory (Hochreiter and Schmidhuber 1997) or using attention-like filters to selectively propagate task-relevant information, such as transformers (Vaswani et al. 2017). Around the same time, deep reward-learning networks were being trained to perform increasingly complex and diverse arrays of tasks (e.g., Mnih et al. 2015). It is notable

From "The Frontal Cortex: Organization, Networks, and Function," edited by Marie T. Banich,
Suzanne N. Haber, and Trevor W. Robbins. 2024. Strüngmann Forum Reports, vol. 34,
Julia R. Lupp, series editor. Cambridge, MA: MIT Press. ISBN 9780262549530

that the evolution of intelligence observed in these network models reflects the iterative addition of each of the canonical computations described above. Parallel feedforward models are able to integrate information effectively; recurrent networks are able to maintain information actively; transformers and long short-term memory rely on selection of feedforward or recurrently maintained representations; and deep RL relies on monitoring to learn and update representations.

This evolution suggests that understanding the mechanisms supporting intelligence in artificial agents may provide a new angle to understanding human intelligence and the role of the frontal lobe, with the hope that some of these mechanisms will be similar to the ones observed in the brain. This development could be useful to gain mechanistic insight at a few levels.

First, computational models may provide insight into the mechanisms and functionality of the frontal lobe. By training computational models on increasingly complex, more "real-world" tasks, we can use the analytical approaches described above to decompose them into underlying computational motifs. Early attempts are already providing new insight into complex cognition: fully recurrent neural networks that are trained on complex context-dependent decision-making tasks show low-dimensional dynamics that are compositionally combined to perform more complex tasks (Yang et al. 2019). One difficulty is that it is often hard to understand how these dynamics emerge from the underlying circuit. In other words, are we simply swapping one complex system for another, slightly less complex system? One potential way to overcome this dilemma is to constrain these models to be low-dimensional (e.g., low rank connectivity) yet still recapitulate the function of more complex models. This approach often leads to more interpretable circuit mechanisms and can reveal computational motifs that align well with previous hand-built models, such as using gain modulation to do context-dependent computations (Dubreuil et al. 2022). This approach perhaps gives us some hope that complex models trained to perform complex behaviors could help us understand how previously known circuit and computational motifs are engaged during real-world behaviors.

Second, understanding AI may provide insight into the canonical computations that are critical for cognition. In other words, studying artificial agents may reveal new canonical computations that we have yet to consider. To a certain extent, such insights have been observed in the application of transformers to large-language models. While theoretical modeling focused on the learning of grammatical structure to generatively produce language, large-language models have demonstrated the power of a simple learning rule, predictive learning, in being able to learn and generate language (Piantadosi 2023). One could imagine similarly surprising insights emerging from AI agents trained to perform complex, real-world behaviors.

Experimental Approaches: Limitations, Opportunities, and Future Directions

Room for Improvement in the Assessment of Function

Goal Selection

In the vast majority of research, participants are either given their goal for a given trial explicitly (e.g., name the color of this word) or are able to infer it from the reward structure of their environment (e.g., it is currently most rewarding to focus on the shape feature). In real life, individuals typically have to set their own goals, including what task to complete and how to complete it. In failing to capture this element of ecology, these studies also fail to capture processes that are commonly impaired in patients with prefrontal lesions; namely, how a person selects their current goal and the subgoals that will help them achieve it (see Table 12.2). Patients with PFC lesions demonstrate substantial task initiation costs, goal neglect, and forms of apathy and avolition that could at least partially reflect an inability to settle on and sufficiently activate an immediate goal. Pathology aside, understanding goal selection can provide better insights into individual variability over development and across individuals in adaptation to the level of "goal scaffolding" within a person's environment (e.g., the extent to which their caretakers provide clear structure for their future aims).

There have been a number of attempts to lend further experimental insight into the process of goal selection, including the classic Multiple Errands Test (Shallice and Burgess 1991). Briefly, patients were sent out on their own to complete a series of errands of varying complexity around an area of London, including purchasing specific items and finding out particular types

Table 12.2 Directions for improvement in existing experimental approaches.

Domain	Common approach	Example novel directions
Goal selection	Explicit and/or well-constrained task goals	Choice of which task to perform when
Planning complexity	Planning over limited number of steps	Larger space of options and potential subgoals
Response complexity	Limited number of discrete and irrevocable actions	Continuous action space, reversible
Value of information	Limited opportunities for and scope of new information	More information-rich tasks and exploratory opportunities
Changes over time	Measures averaged over the course of a single session	Analyze temporal dynamics within/across many sessions
Naturalistic measures	Tasks performed in the lab	Tasks and other measures (EMA, physio, mobile EEG) measured out "in the wild"

From "The Frontal Cortex: Organization, Networks, and Function," edited by Marie T. Banich,
Suzanne N. Haber, and Trevor W. Robbins. 2024. Strüngmann Forum Reports, vol. 34,
Julia R. Lupp, series editor. Cambridge, MA: MIT Press. ISBN 9780262549530

of information from those items, such as the exchange rate of the French franc on the previous day. The errands required following specific rules to achieve those goals, such as not entering a shop without buying something. Compared to matched controls, the frontal lesion patients broke the rules more often and performed the task less efficiently, including exhibiting problems with the selection and implementation of subgoals. Qualitatively, a patient's behavior was unlike any of the controls. For example, one patient picked the wrong newspaper, did not pay for the item, and ended up being chased by the shopkeeper. Another focused on buying soap that she preferred rather than adhering to the instructed goal of obtaining the cheapest available soap. In essence, these patients appeared to exhibit difficulty both in adequately selecting and carrying out specific goals, as well as in implementing behaviors that were most appropriate to achieve those goals (i.e., rule breaking).

More recent experiments have examined much simpler forms of goal selection within the laboratory, by allowing, for instance, participants to choose freely which of a limited set of tasks to perform, and for how long, based on factors like expected reward and difficulty (Arrington and Logan 2004; Gilzenrat et al. 2010; Orr and Banich 2014; Parro et al. 2018; Westbrook et al. 2013). Some of these tasks have provided evidence of prefrontal involvement in task choice (e.g., Orr and Banich 2014; Westbrook et al. 2019; Wisniewski et al. 2015). Other work has examined another key dimension of goal selection; not *which* task to perform but *when* to perform it. For instance, Le Bouc and Pessiglione (2022) had participants perform laboratory choice tasks that assessed the extent to which they preferred exerting effortful tasks later rather than sooner, and then showed that these task-based estimates predicted how long participants would wait before returning a set of forms they had been asked to complete and return any time within the next month. These experiments offer instructive examples of studying the various dimensions of goal selection within a controlled environment. Nonetheless, they fall substantially short of capturing the complexity of real-world goal selection, as exemplified in the errands task above.

Planning Complexity

In line with the above-discussed desire to understand goal selection and changes in PFC function across time, tasks that require multistep planning (as would be required e.g., during cooking or playing Atari games), might be a particularly useful tool for studying PFC function. Planning tasks often require internal simulation before a choice is made, thus tapping into one of the main adaptive functions of PFC described above. Planning tasks can also incorporate a reward-learning process, which then opens a window into the relative roles of forward and backward simulations for planning and learning processes, as well as the goal-oriented cognitive map over which planning occurs (Mattar and Daw 2018). It might be particularly instructive to investigate forms of repeated planning that can be optimized over time, providing greater insight into the process by which subgoals are learned. In addition to the insights these tasks provide into planning itself, tasks like these have the added benefit that they involve a comparatively high number of options or action sequences (e.g., Eldar et al. 2020; Huys et al. 2012; Kurth-Nelson et al. 2016), which is another desirable property discussed further below.

Response Complexity

In general, behavioral tasks are conceived to be as simple as possible to expedite training for participants and make the space of analyses/interpretations as narrow and tractable as possible for the experimenter. Tasks that are too simple, however, might not engage prefrontal mechanisms, thereby obscuring anatomical and functional segregations, making it difficult to discriminate between computational models of prefrontal function. For instance, tasks that force a choice between two responses face a challenge disentangling between selection of one response and inhibition of the alternate response. Prefrontal functions are critical to manage real-life environments that feature high-dimensional, uncertain, changing and open-ended situations as well as continuous and often reversible behaviors. Investigating prefrontal functions certainly requires a consideration of behavioral paradigms that capture these complexities as much as possible.

The Value of Information

Another higher-level process that is believed to be supported by the frontal lobe is exploration (Badre et al. 2012; Domenech et al. 2020; Monosov and Rushworth 2022), including tracking properties of the environment that give rise to the antecedent experience of curiosity. This process serves to identify a conceptual space of potentially useful information or behaviors that *might* be relevant in the current context or potentially useful in the future. Acquiring knowledge about the environment to learn proper internal world models is central to efficiently fulfilling the ever-changing needs of the organism (Koechlin and Wang, this volume). Information-seeking is thus believed to constitute a primary drive of behavior and is potentially separate from reward-seeking. How the PFC arbitrates between reward-seeking and information-seeking motives, and the extent to which information-seeking serves to maximize expected future outcomes and/or minimize aversive uncertainty, awaits further research, both computational and empirical (cf. Cockburn et al. 2022; van Lieshout et al. 2019, 2021a, b). Doing so will benefit from novel experimental designs that incorporate a wider range of potential future states and varying motives for seeking out or avoiding those states, under varying levels of known or unknown uncertainty.

From "The Frontal Cortex: Organization, Networks, and Function," edited by Marie T. Banich,
Suzanne N. Haber, and Trevor W. Robbins. 2024. Strüngmann Forum Reports, vol. 34,
Julia R. Lupp, series editor. Cambridge, MA: MIT Press. ISBN 9780262549530

Changes in PFC Involvement over Time

Research into the broad set of functions laid out above often examines behavior and neural activity averaged over the course of an experiment (e.g., trials within a session). In doing so, these studies miss changes that occur over periods of time within a session that could offer critical insights into the drivers and dynamics of PFC function. For instance, over the course of a single experiment, attention, effort, and control demands may vary (e.g., due to boredom, mind-wandering, fatigue, practice, and/or fluctuations in mood); learning is likely to occur (shaping changes in task-relevant representations); and participants may shift between strategies for performing the task. These factors all raise the potential for increased measurement noise. More importantly, they also represent missed opportunities for understanding these functions at a finer grain (e.g., mechanisms of plasticity, distractor interference, influences of motivation and affect on controlled processes).

These dynamic changes have raised particularly acute concerns about the extensive training that occurs prior to nonhuman animals performing such tasks. This limitation also introduces opportunities, both for beginning to examine performance over the course of this extended training regime (e.g., Masís et al. 2023) and for examining human parallels to such extensive levels of training (e.g., Balci et al. 2010; Blain et al. 2016). For example, a recent study by Miller et al. (2022) carried out extended testing in human participants over the course of three months on both a working memory and serial reaction time task. Working memory performance improved throughout this time window, and significant evolution was seen in delay period activity patterns in the frontal lobe. More generally, though, these opportunities should be more regularly exploited over longer timescales, within both humans and other animal models, by studying how cognitive functions, neural anatomy, and physiology vary over the course of multiple experimental sessions, days, weeks, or months apart (e.g., Allen et al. 2022; Naselaris et al. 2021; Poldrack et al. 2015).

Measures of Naturalistic Behavior

As researchers, we are interested in understanding and predicting behavior outside the lab. Doing so inevitably will involve considering the greater diversity of environmental contexts that people experience. The real world is distracting, noisy, and variable in terms of resources. More naturalistic assessments may be helpful for understanding how differences in these and many other factors may affect PFC functioning. For example, having participants complete tasks in their homes or on their phones may provide a better understanding of real-world performance. Ecological momentary assessments (EMA), which prompt participants to answer questions about their experiences at that particular moment (e.g., their current goal, emotional state, or context), may provide insight into everyday behavior and variability (e.g., Hofmann et al. 2012b). One might also obtain measures relevant to function from passively collected data, such as global positioning satellite locations or accelerometry measures from wearable devices (e.g., Heller et al. 2020), which would help reduce participant burden and remove biases that may arise with self-report measures.

Advances have been made in the tools researchers have at their disposal to measure PFC activity out in the world, such as functional near infrared spectroscopy (fNIRS) or mobile EEG. With these measures, neural mechanisms, which until now have exclusively been investigated in laboratory environments, become translated to natural situations with the advantage that concepts about the neural implementation of PFC functions can be put to the test in natural environments. Such endeavors will significantly broaden the validity of the already established concepts about prefrontal neural processes, ultimately leading to a more holistic understanding of PFC function (see Table 12.2).

Leveraging Recent Advances in Data Analysis

Deep Convolutional Neural Networks/Deep Q Networks

One promising approach for probing the nature of the representations found in the PFC involves the use of network models imported verbatim or with small modifications from the AI literature. These models can either be pre-trained to perform specific tasks, such as object recognition (Kriegeskorte 2015), or trained from scratch to perform specific tasks, such as learning to play particular Atari games (Mnih et al. 2015). Although these models are very different from the architecture of the brain, both in terms of their physical structure and the rules used for modifying plasticity within them, they have been successful in revealing patterns of activity in their layers that seem to correspond broadly to patterns of activity in the brain (at the level of single neurons or populations) and fMRI activity, when applied to activity measured while animals or humans are performing the same tasks on which the network itself has been trained (Cross et al. 2021; Iigaya et al. 2023; Kriegeskorte 2015; Yamins et al. 2014). Though these approaches have been mostly used to date to illuminate representations in the ventral and dorsal visual stream as opposed to the PFC, it is likely that models incorporating more complexity, such as recurrency and/or multinetwork structure, may prove useful in explaining patterns of activity in the PFC as well (Perich and Rajan 2020). One important way to leverage these models is to explore whether some variants on their architecture can better account for neuronal activity than others. Furthermore, inducing lesions in those models and seeing to what extent particular components of the model are critical for behavior might also serve as a basis for refining hypotheses about causality regarding particular prefrontal areas, which could then be tested in future causal perturbation experiments, such as with inactivations, optogenetic or chemogenetic manipulations in animals, and/or transcranial magnetic stimulation.

Another approach would be to use deep convolutional neural networks—which have been optimized for the structure of neurophysiological (e.g., EEG) data (Lawhern et al. 2018)—to analyze such data in a way that makes it possible to delineate potential novel features that had been potentially overlooked by theory-driven approaches (Vahid et al. 2020). Through the use of explainable AI methods, the novel features identified could then be integrated into existing conceptual frameworks on the cognitive processes being examined in the study at hand. Moreover, other methods, such as generative adversarial networks (Goodfellow et al. 2014), may provide valuable insights into the neurophysiological principles underlying cognitive functions supported by the PFC. For example, these networks have been used to show that neurophysiological principles of two opposing instances of cognitive control processes or antagonistic behaviors can be transferred to each other (Vahid et al. 2022). Since such deep learning procedures are able to capture nonlinear interdependencies, these approaches may be well suited to examine the interrelation of neural principles that are associated with the above-mentioned canonical computations.

Combining across Data from Multiple Tasks

As outlined above, it is likely the interconnection between different canonical computations and the relative weighting of the computations are important to understand in PFC function. It is, therefore, important to abstract from the level of specific tasks and analyze neural data in a more overarching, task-invariant way. This approach has particularly been lacking within analyses of neural time series data. Principal component analysis (PCA) and independent component analysis (ICA) have been used to extract neurophysiological components but are optimized for two-dimensional data (e.g., covering spatial and temporal information of time series data). However, data from typical experiments with concomitant data recordings (e.g., EEG) can yield more dimensions: time, space, frequency, trial, condition, participant, and group (Cong et al. 2015). These dimensions can mathematically be described as tensors. Applying methods optimized for two-dimensional data (i.e., PCA and ICA) in the face of such data is only possible by reducing data dimensionality (e.g., by concatenating or stacking the data). This, however, leads to an inevitable loss of information (Cong et al. 2015).

Tensor decomposition techniques can capture additional dimensions of information contained in neural time series data (Cong et al. 2015). Through these techniques, factors such as "tasks" can directly be modeled in the data analysis, allowing one to look at possible distinct and common neural profiles across tasks. This approach may provide a necessary step toward a thorough examination of neural principles across tasks and probable distinct or common profiles of canonical computations mediated by the PFC. Crucially, this method also overcomes another important shortcoming of most strategies used in the analysis of neural time series data: the reliance on averaged parameters of neural activity, which tacitly assumes that neural processes do not change across time spent engaging in various aspects of canonical computations mediated by the PFC. Through tensor decomposition methods, it is possible to better model moment-to-moment variations and changes over time in the neural activity profile without losing the information of other relevant dimensions in neural time series data.

Data-Driven Identification of Functional Primitives/Common Motifs across Tasks

Classically, our understanding of the neural mechanisms underlying cognition has been "top-down". We use theoretical concepts to generate hypotheses which, in turn, drive experimental design and data analysis. Recent work has begun to take a more data-driven approach to identify processes of interest. For example, combining cutting-edge factorization and dimensionality-reduction techniques has allowed us to begin to decompose naturalistic behavior into a sequence of action primitives, referred to as "behavioral motifs." The transition between behavioral motifs has been related to striatal activity, providing a novel perspective for understanding the function of striatum (Markowitz et al. 2018, 2023; Wiltschko et al. 2015). Similarly, the spatiotemporal pattern of cortex-wide neural activity can be decomposed into a set of ~15 dynamic "neural motifs" (MacDowell and Buschman 2020). These motifs repeat over time, across tasks, and between individuals, suggesting they provide a canonical basis set of underlying patterns of neuronal firing (i.e., primitives) that aid in understanding the dynamics of neural activity across the cortex.

Can one take a similar approach to prefrontal function? To do so, it would be desirable to define PFC-dependent functional primitives objectively, in a data-driven way. Quantitative cognitive ontology is becoming possible with the help of large-scale population experiments and machine-learning aided data analysis (Eisenberg et al. 2019). While data-driven approaches will likely provide support for the theoretically motivated hypothesized cognitive functions, the hope is that they may also identify novel mechanisms that have not previously been considered.

Dynamical Systems, Subspaces, and Neural Geometry

A relatively new approach to describing neural representations is at the population level, providing new insight into the dynamics and representations of the frontal lobe. Large-scale recording approaches have allowed researchers to track the activity of an ever-increasing number of neurons. One can visualize patterns of activity across the entire population of neurons as a point in an N-dimensional space (where N is the number of recorded neurons and thus, very high). Recent work has begun to understand how the geometry of these representations may enable generalization and compositionality (Bernardi et al. 2020; Fu et al. 2022; Panichello and Buschman 2021; Weber et al. 2023).

Dimensionality-reduction techniques and classifiers can be used to identify low-dimensional subspaces within the high-dimensional neural space that encodes task-relevant variables. Understanding how these subspaces relate to cognition is a rapidly emerging field. For example, computational modeling suggests "learning-to-learn" is facilitated by creating a subspace within PFC that is shared across a series of tasks (Goudar et al. 2023).

The dynamics of neural activity can be quantified by considering the trajectory of neural activity in state space (Shenoy et al. 2013). This approach has provided insight into how neural representations evolve over time. For example, sensory representations have been found to rotate over time, eventually forming a short-term memory representation of the stimulus input (Libby and Buschman 2021). This rotation allows both sensory and memory information to be represented in independent subspaces. Importantly, because these subspaces are orthogonal to one another, sensory and memory representations do not interfere with one another. In the context of sequence learning, the coexistence of sensory and memory representations may be important for associative learning. More broadly, such rotations may be important for reducing interference in (a) working memory (Panichello and Buschman 2021), (b) between representations of targets of attention (requiring selective enhancement) and distractors (requiring selective suppression) (Ritz and Shenhav 2024), and (c) between tasks (Weber et al. 2023).

Finally, subspaces may allow for the routing of information between brain regions. Simultaneous recordings within visual cortex identified a subspace within V1 that "communicated" with V2: changing the neural activity within this subspace influenced downstream activity, while changing the neural activity outside of this subspace did not (Seitzman et al. 2019). Larger-scale recordings have shown that these communication subspaces are not one-to-one, but rather extend to broader networks of regions (MacDowell et al. 2023). This arrangement then may provide an ideal mechanism for cognitive control. Changing how information is represented within a given brain region could change how that information is propagated to other regions. For example, representing information in "private" dimensions (that are not communicated) could keep information local, while transforming that representation into a "shared" subspace could broadcast that information to other brain region(s).

Conclusions and Open Questions

Historically, research into the function of PFC has been driven by numerous conflicting theoretical accounts and insufficient and/or inconsistent evidence to constrain or adjudicate among them. We have sought to cut through these conflicts by offering an integrative perspective on the common computations that

From "The Frontal Cortex: Organization, Networks, and Function," edited by Marie T. Banich,
Suzanne N. Haber, and Trevor W. Robbins. 2024. Strüngmann Forum Reports, vol. 34,
Julia R. Lupp, series editor. Cambridge, MA: MIT Press. ISBN 9780262549530

underpin previous findings within PFC, including both underlying patterns of neural activity and manifestations of damage to its regions. The four canonical computations that we have highlighted—integration, maintenance, selection, and monitoring/evaluation—offer a parsimonious account of the interlocking computations that are both necessary for goal-directed behavior and potentially sufficient for explaining the array of observations just noted.

The account we have provided is, of course, incomplete and in many ways demands further iterative refinement and revision. As one example, in accounting for the broad set of functions that we outlined at the start of this chapter (e.g., planning, flexibility, and active maintenance), we offered a set of common underlying computations that largely accorded with (and deliberately integrated over) ones that have been previously proposed across different literatures. Future work should seek to identify new computations to augment or even replace those we offered.

There are also a number of questions that we were unable to address but which will be critical for providing a comprehensive account of PFC function. How do PFC computations vary in the timing of their engagement within a given task trial (manifesting, e.g., as difference in proactive versus reactive control)? How are symbolic representations (e.g., language) incorporated into core computations (integration, selection, maintenance, and monitoring). and how does this emerge over evolutionary development (e.g., with human participants able to learn and adapt flexibly based on verbal instructions alone)? To what extent do these computations give rise to key elements of social cognition (e.g., mentalizing, perspective-taking), and in what ways are they supported by other core functions within or outside of PFC? In what ways is engagement of PFC functions experienced by the organism as effortful, and to what extent do these functions each depend on motivational input to carry out versus continuing automatically in the absence of motivation (Shenhav et al. 2017; Westbrook and Braver 2015)? Finally, how are PFC functions facilitated by and/or interfered with as mood and affective states vary (Kenwood et al. 2022; Pizzagalli and Roberts 2022)? Answers may provide important clues into the role of PFC dysfunction versus neuromodulation in generating versus alleviating symptoms of certain psychiatric disorders, such as major depression (see Rowe et al., this volume).

From "The Frontal Cortex: Organization, Networks, and Function," edited by Marie T. Banich,
Suzanne N. Haber, and Trevor W. Robbins. 2024. Strüngmann Forum Reports, vol. 34,
Julia R. Lupp, series editor. Cambridge, MA: MIT Press. ISBN 9780262549530