Neurocomputational Models of Task Representation

Michael C. Freund*[1,2] and Todd S. Braver[2,3]

[1]Department of Cognitive, Linguistic, and Psychological Sciences, Brown University
[2]Department of Psychological & Brain Sciences, Washington University in St. Louis
[3]Department of Radiology, Washington University in St. Louis

December 2023

Author Note

This manuscript appears as a chapter in The Sage Handbook of Cognitive and Systems Neuroscience. The content was largely abridged from a 2021 subject-matter exam in partial completion of requirements for a PhD in Psychology at Washington University in St. Louis. Drs. Wouter Kool and Jeffery Zachs provided constructive comments on the original version of this manuscript.

*michael_freund@brown.edu

Contents

1 Introduction

Human cognition is purposive. We can set our minds toward attaining a particular desired goal, ignoring distractions and overriding counterproductive habits. Yet, it is also multipurpose. We are not permanently bound to pursue a singular goal, or even a small number. As drives and circumstances change, we can flexibly set our minds to attaining more suitable ones. How does the brain achieve this "setting" and "resetting" — this capacity for cognitive control? The promise of understanding this fundamental question has vigorously motivated many cognitive neuroscientists over the past half-century and continues to do so today.

Increasingly, cognitive neuroscientists are focusing on studying representations that underlie cognitive control, formalizing their ideas within neurocomputational models (Botvinick and Cohen, 2014). Two distinct models of task representations have been highly influential: Guided Activation (Miller and Cohen, 2001), and Adaptive Coding (Duncan, 2001) through Random Mixed Selectivity (Rigotti et al., 2013). While these models have several commonalities — both posit that the prefrontal cortex (PFC) drives controlled behavior through task representations — they differ greatly in the representations they posit, and more generally reflect discrepant philosophies of understanding brain function.

In this chapter, we provide an overview of cognitive research into cognitive control, focusing on the role of task representations. A detailed examination of the Guided Activation and Adaptive Coding or Random Mixed Selectivity models is provided, in which models are evaluated in relation to the extant experimental literature and compared with respect to broader issues regarding task representations.

1.1 Definitions and Scope

1.1.1 Representation and task set

The term "representation" is often used, but also frequently debated in cognitive neuroscience. In general, it refers to a neural correlate, that is, a relationship observed between modulations in brain activity and manipulated variables, such as an experimental condition (e.g., a stimulus feature). More philosophical uses reflect the stance that a neural population can be described as implementing a transform, in which certain dimensions of input information are "extracted" or emphasized, while other dimensions are "ignored" or deemphasized (deCharms and Zador, 2000). In this sense, representation implies causal function: *to signal* the extracted information to whatever populations lie downstream, and ultimately *to implement* cognitive processes or behavior. Even stronger conceptualizations of representations exist, too, that additionally imply stability over time — for example, a pattern of synaptic weights, stored long-term, that has causal impacts when it is reactivated (Wood and Grafman, 2003; cf. Barack and Krakauer, 2021). Similar to others (Ebitz and Hayden, 2021), when discussing empirical findings in this review, we will use "representation" (or "encode", or "code") in its more general sense (i.e., as a correlation), but when discussing theoretical objects, such as aspects of computational models, we will use it in its narrower sense (i.e., as a causal transform). The context should be appropriate to disambiguate.

A central construct in cognitive control research is a "task set". A task set refers to the particular configuration of cognitive resources that enables one to perform the task at hand (Monsell, 2003, 2017; Sakai, 2008). As a construct, of course, task sets are not directly observable. But, because

they are posited to involve "cognitive resources", aspects of them should be observable in behavior and in brain activity. In terms of information processing, task sets include information used to perform the task, which usually pertains to the stimuli, responses, mappings among them — or *rules* — the required cognitive operations, and how all of these elements relate to one another temporally. In terms of the brain, a task set can be reflected at multiple scales: from a large-scale network configuration that supports task-appropriate flow of information, to more focal coding of task elements within a brain area or neural population. This review focuses on a vital element of a task set, *task representations*, which are generally thought to be contributed by PFC circuitry and to orchestrate other components of the task set. Recent reviews have used the term "control representation" (Badre et al., 2021), in an equivalent manner to how "task representation" is used here. We prefer "task representation" to exclude representations that might be more relevant to other control processes, such as those associated with conflict or error monitoring, and gating or updating of working memory (Alexander and Brown, 2011; Brown and Braver, 2005; O'Reilly and Frank, 2006).

1.1.2 Contextual control and mid-lateral PF

The scope of this review was constrained to focus primarily on what has been termed "contextual control", or control over thoughts and actions within the present moment according to some internally represented context, such as a rule (Badre and Nee, 2018). One distinguishing feature of contextual control is its timescale, referring to processes and behaviors that span, from perception to action, several seconds, rather than minutes or hours (cf. schematic and episodic control; Badre and Nee, 2018). Another feature of contextual control is that it involves contexts or rules that generalize over multiple stimulus–action mappings, and thus are to some degree abstract. While the particular degree of abstraction has featured in several influential accounts of PFC organization (Badre, 2008; Koechlin et al., 2003), we do not focus on this issue here, and instead primarily review research that concerns relatively intermediate levels of abstraction (Badre and Nee, 2018 for review). Accordingly, we also focus primarily on brain areas within *mid-lateral PFC* (areas 46 and 9/46d of Petrides and Pandya, 2001), referred to here as dorsolateral prefrontal cortex (DLPFC), as much research shows these areas are critically involved in implementing contextual control.

Additionally, while not an explicit guideline, the tasks used in research we review mainly involve executing actions that are typically eye, hand, or mouth movements, in response to stimuli that are typically visually presented. Undoubtedly, these sensorimotor interfaces are vital to the evolutionary strategy of the primate. But this focus leaves underexplored a wide range of task sets that involve, for instance, more internal or ruminative functions (i.e., in the absence of impending action), or more complex stimulus–response sets.

2 Guided Activation

Miller and Cohen (2001) synthesized foundational neuropsychological, neuroscientific, and psychological research into a coherent "guided activation" theory of cognitive control and prefrontal function. Guided Activation theory advanced two broad hypotheses regarding how PFC exerts control. (1) Neural populations in PFC represent internal behavioral and cognitive goals in their distributed patterns of activity. (2) These PFC goal representations bias the competition of posterior cortical representations in favor of representations that allow one to achieve the internal

goal.

A potent metaphor was used to illustrate this model, of railroad tracks and operators. The context of this metaphor was the notion that, generally, the goal of a cognitive process is to transform certain inputs (e.g., sensory stimuli) into appropriate outputs (e.g., an action) through input–output pathways. In the metaphor, these pathways are railroad tracks, laid down over time via practice and associative learning. Learning to read, for example, would eventually build a track from the word "RED" to its corresponding verbal response. As in the case of reading, which is both highly practiced and highly specified by textual stimuli, cognitive processes can often unfold relatively automatically, without top-down intervention from control systems. In these cases, the "RED" train would run freely and arrive on time. However, sometimes the path of multiple, simultaneously running trains attempt to navigate the same tracks — for example, when naming the color of a color-word stimulus (Stroop, 1935). In these cases, an external switch operator, PFC, is required to step in and manage traffic flow.

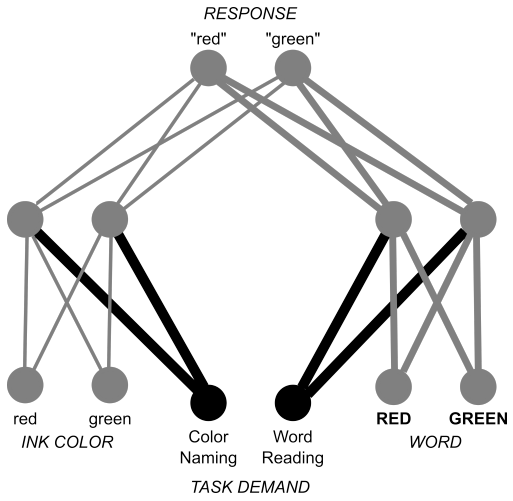

Figure 1: **Guided Activation theory, formalized within the feed-forward neural network model of Cohen et al (1990).** Stimulus features (ink color, word) are fed forward to a hidden layer, which transforms these inputs into responses. Relative to ink-color inputs, word inputs are more strongly connected to responses, reflecting the difference in automaticity of word-reading and color-naming tasks (illustrated by the width of lines). Control is implemented via the "task demand" units, which project to the hidden layer (black nodes and lines). In the color-naming task, bias from top-down control is needed to redirect activity flow along the weaker (less automatic) color-naming pathway.

Guided Activation theory was based heavily on Cohen's (1990) neural network model of the color-word Stroop task (Figure 1). "Task Demand" units in Cohen et al. (1990) correspond to populations of prefrontal neurons, which represent the currently active goal ("Color Naming" units; Figure 1). These units selectively excite intermediate units of the color-naming pathway, which correspond to more posterior sensorimotor or semantic representations of the stimulus hue and associated response. This "bias" in favor of the color-naming pathway temporarily allows activation of the task-relevant color response to readily accumulate in the output layer, regardless of co-activation of potentially competing responses driven by the word-reading pathway. Such bias occurs in a *top-down* manner, as it reflects the internal goals, rather than external or *bottom-up* goals that may

be activated transiently by sensory input.

The Guided Activation model was closely related to the Biased Competition model of visual attention (Desimone and Duncan, 1995). In biased competition, similar competitive interactions among visual representations were proposed to occur when visual attention is guided to an object or feature. Resolution of such competition was hypothesized to filter or select representations for subsequent processing along dorsal and ventral visual streams. While this model proposed that many such competitive interactions can occur in parallel and in a bottom-up manner throughout the visual system, the ultimate source of top-down attention was proposed to be PFC. Guided Activation broadened this idea, as a proposal for how such top-down attention may be exerted generally on representations of posterior cortex. In this way, the Guided Activation model was an attempt to model cognitive processing, from lower-level sensory attention to higher-level cognitive control functions, using domain-general computational principles.

Guided Activation emphasized the distinction between "controller" systems of PFC, and "controlled" systems, such as more posterior sensory and motor-related circuitry. PFC was assumed to be a dedicated controller, which biases distal representations that are "responsible for actually performing the task" (Miller and Cohen, 2001). From this view, the role of PFC is *modulatory*: an auxiliary circuit, external to the stimulus–response pathway that influences downstream processing in the service of appropriately mapping stimuli to responses. In the railroad metaphor, PFC is the switch operator that redirects the train along extant tracks (Figure 2, left). This modulatory view can be contrasted with a *transmissive* view (Badre et al., 2021; Miller and Cohen, 2001), in which PFC is an integral part of the stimulus–response pathway (e.g., a set of stations through which all trains pass; Figure 2, right). The modulatory role of PFC is a key feature of the Guided Activation model, reflecting a broader view on the large-scale organization of the brain and the position of PFC within this circuitry.

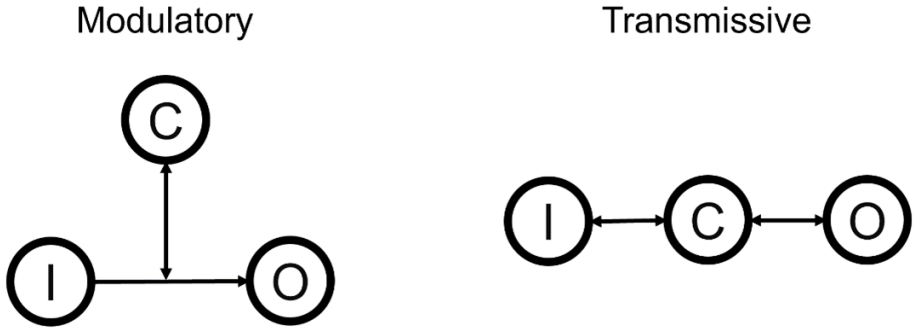

Figure 2: **Modulatory versus transmissive architectures of cognitive control and PFC.** I: input (e.g., stimulus); O: output (e.g., response); C: control.

2.1 Key Computational Mechanisms

To guide activation, the prefrontal cortex must represent appropriate mappings between inputs and outputs. But, given our expansive behavioral repertoire, yet limited cortical tissue, it is not possible to have all potential input–output mappings explicitly represented in the PFC. What kind of organizational scheme might underlie these internal goal representations? Miller and Cohen

(2001) did not directly address this issue, except to speculate that plasticity likely plays a major role. Nevertheless, one suggestion is clearly implied by the format of PFC representations within their model (Cohen et al., 1990) — namely, that PFC represents behavioral goals within an abstract format. Use of an abstract scheme could provide considerable compression: instead of every input–output mapping, only *classes* of mappings would need to be represented.

Although the notion that PFC represents newly acquired abstract rules to guide behavior was not new (Cohen et al. 1990; Passingham, 1993; Wise et al., 1996), some monkey neurophysiological studies at the turn of the 21st century provided clearcut empirical evidence for this idea. Critically, these studies were carefully designed such that neural responses could be verified as "abstract" —– that is, not selective to certain stimulus or response features, but instead to a rule that defines how to respond to different stimuli. Essentially, this was done by equating as much as possible stimuli and responses across two rule conditions of a task. In one exemplar study of this type (White and Wise, 1999), monkeys were required to manually respond when a subtle visual target (a small change in luminance) appeared at one of four locations. The target was undetectable unless its location was foveated. The relevant location was instructed via pre-cue, according to one of two rules: a spatial cue, presented at the target location, or a shape cue, arbitrarily associated to the target location. Because the sensory and requisite responses of the trials were equated across the spatial and associative rules, this design enabled abstract rule-specific activity (i.e., independent of stimulus or response feature) to be examined. Approximately one-third to half of the lateral PFC neurons that exhibited task-related modulations were selective to one rule or the other (approximately equally balanced across rules). Subsequently, abstract coding in PFC was more extensively explored in monkey within a series of seminal studies (Freedman et al., 2001; Nieder et al., 2002; Wallis et al., 2001; cf. Miller et al., 2003). Convergent findings were contemporaneously reported in humans, by capitalizing on advances in fMRI event-related designs (Bunge et al., 2003; MacDonald et al., 2000). Collectively, these studies definitively illustrated that lateral PFC has the capacity to flexibly encode verifiably abstract, behavior-guiding rules.

Building on these findings, an extension to the Guided Activation model formalized abstraction within a computational model of PFC rule representations (Rougier et al., 2005). The model was trained to perform a variant of the Wisconsin Card Sorting Task (WCST), which required multi-dimensional objects to be matched along only one dimension (four dimensions were used: color, shape, number, texture; for a simplified example of the task, skip ahead to Figure 3A). After training, the PFC layer was indeed shown to provide bias to the response layer through representations of the abstract stimulus dimensions — that is, the PFC units were tuned to general dimensions (e.g., "shape") rather than individual features of a dimension (e.g., "circle"). Because these representations were abstract, the network could perform the various tasks on novel (untrained) stimuli.

This model was able to learn abstract representations because of a combination of features. The first was an adaptive gating mechanism, modeled after the hypothesized function of striatal–PFC circuitry (Braver and Cohen, 2000; O'Reilly and Frank, 2006). The second was blocked training on a diversity of features. The adaptive gate enabled PFC representations to be stabilized — that is, robustly maintained in an active state — throughout a training block, in which only one dimension was relevant. Within these blocks, training on a diversity of features allowed for the model to learn which patterns of stimulus inputs were (ir)relevant for the current dimension. When a training block ended and a new began, the adaptive gating mechanism destabilized PFC activity

(enabling PFC activity to encode other activity patterns), discouraging learning of features across dimensions. Thus, these basic mechanisms were found to be sufficient for learning abstract task representations.

2.2 Empirical support

2.2.1 Representation of internal goals

If PFC indeed represents internal goals, then neural population activity in PFC must reflect the internal behavioral context that the subject is using, whatever this context may be (e.g., an abstract rule, a conditional association between stimuli and responses, a strategy). This hypothesis makes the clear prediction that, at least, task-relevant information should be emphasized relative to task-irrelevant information within prefrontal activity.

Indeed, evidence in monkeys supports this hypothesis. Of recorded monkey prefrontal neurons, a considerable proportion show task-related modulations (Rainer et al., 1998; Wallis et al., 2001). As reviewed above, many of these also show selectivity to behaviorally relevant rules, categories, stimulus features, actions, conditional associations between stimuli and responses, and conjunctions of various task-relevant aspects. By contrast, while task-irrelevant information is not absent (Chen et al., 2001; Genovesio et al., 2009, 2011; Mante et al., 2013), it is often less emphasized within the code, at least in terms of number of neurons showing significant modulation to task-irrelevant dimensions.

Selectivity of LPFC to task-relevant information is also supported by many human neuroimaging studies. The fact that human LPFC is recruited more strongly when demands are placed on preparing for a task on the basis of a rule (Bunge et al., 2003; MacDonald et al., 2000), suggests, albeit indirectly, that human LPFC encodes task-relevant rule information. More direct evidence for the prioritization of task-relevant information has been provided by multivoxel pattern analysis (MVPA) techniques, which essentially enable differences in the representation of task conditions to be examined with greater sensitivity, even if they are not associated with overall, region-wide differences in activation. In particular, MVPA techniques have been critical for studying the neural representations of conditions that have similarly high levels of control demand, such as abstract rules or attentional target features. Indeed, with these methods, human LPFC has been confirmed to represent a diversity of task-relevant information, including abstract task rules, sensory-related information, and target actions or response-related information (Woolgar et al., 2016). During task switching tasks, for example, the task rule was shown to be robustly encoded in LPFC in preparation for the trial (Etzel et al., 2016; Haynes et al., 2007). In a delayed match-to-sample task, stimulus-specific information was not detected within LPFC, but currently active task rule was (Riggall and Postle, 2012). More recently, in a color-word Stroop task, the target (task-relevant) response was shown to be encoded more strongly than the distracting (task-irrelevant) action in DLPFC, and was a robust predictor of the speed with which subjects resolved Stroop interference (Freund et al., 2021).

2.2.2 Top-down bias

If controlled behavior indeed relies on prefrontal rule or goal representations to enhance posterior representations in a top-down manner, then changes in such PFC representations should

correspond to changes in posterior representations, which should in turn be linked to changes in behavior (Passingham and Wise, 2012). What evidence would be consistent with such a circuit? This question has been most extensively studied within the domain of visual attention.

As a prerequisite for top-down bias, there should be goal-related enhancements in posterior representations. Indeed, directing attention to a visual feature, object, or location is known to modulate coding in visual cortex in ways that selectively enhance readout of goal-relevant information (Maunsell, 2015; Maunsell and Treue, 2006; Reynolds and Chelazzi, 2004; cf. Egner and Hirsch, 2005; Yeung et al., 2006). Other prerequisites are a division of labor and temporal precedence. Goal coding should be prioritized in prefrontal cortex relative to other brain regions, and should not only emerge prior to the response, but prior to goal coding elsewhere. Several studies employing simultaneous multi-area recording have reported the emergence of rules or target features in frontal areas, including DLPFC and frontal eye fields (FEF), before posterior areas such as inferotemporal (IT) cortex, V4, and MT (Buschman and Miller, 2007; Muhammad et al., 2006; Siegel et al., 2015; Zhou and Desimone, 2011). Collectively, these studies suggest the plausibility of lateral prefrontal areas functioning as a source of top-down enhancements.

But do these rapidly developing goal representations actually drive attentional enhancements? In one clever study, monkey lateral frontal cortex (areas 8, 9, 46, 45, 12) was unilaterally lesioned (Rossi et al., 2009; reviewed in Passingham and Wise, 2012). By additionally transecting the corpus callosum, visual stimuli could be delivered, putatively selectively, to one of two hemispheres: a contralesional hemisphere, which was subject to top-down control of lateral frontal cortex, and an ipsilesional hemisphere, which was not. This provided a within-subject control condition to examine the effects of the lesion. Lesioned monkeys displayed selective but striking deficits in a visual search task: severe impairments in top-down search were observed in the ipsilesional condition, but primarily when the target cue switched frequently across trials. Other supporting evidence involves stimulation paradigms. For example, microstimulation of monkey FEF, below the threshold for inducing saccades, enhanced V4 excitability, and did so in a retinotopically coordinated manner, such that stimulation of neighboring receptive fields induced suppression of V4 neurons (Moore and Armstrong, 2003). Similar downstream enhancements have been found with concurrent TMS and scalp-EEG studies in humans (Morishima et al., 2009; Veniero et al., 2021). Thus, at least in the case of more caudal regions of PFC and the visual domain, it seems likely that prefrontal bias of sensory coding reflects a mechanism of top-down attention. Further investigation is still needed, however, to support stronger claims regarding the generality of top-down biasing mechanisms in lateral PFC. For example, a key open question is whether the caudal lateral PFC goal representations described above are themselves subject to a similar form of top-down enhancements from more rostral lateral PFC representations (e.g., of rules).

2.2.3 A modulatory PFC?

An important aspect of the Guided Activation model is that it proposes PFC-driven control occurs as a top-down modulatory, or biasing, mechanism (Miller and Cohen, 2001). The evidence reviewed above concerning caudal lateral PFC feedback to visual cortex during visual attention, certainly aligns with such a modulatory view. However, lateral PFC also has extensive feedforward connections to premotor cortices, and thus it seems just as likely that control in these tasks could have been mediated by "transmission" forward. One ostensibly straightforward way to reason about this question is through the hypothetical impact of lesions. If lateral PFC could generally

be described as a modulator, then lesions to PFC should not disrupt the lower-order pathways mapping perception to action. Merely the ability to direct this mapping according to internal goals would be impaired. Behavior, then, would simply follow the most prepotent option.

Some neuropsychological evidence is consistent with a modulatory view. As Miller and Cohen (2001) argued, the classic symptoms of "frontal syndrome" suggest a modulatory PFC. This symptom profile is associated with severe disruptions in everyday tasks that require complex sequences of behaviors to be executed (e.g., shopping for food): some components of the behavior are executed, but out of order, or at the wrong times (Stuss and Benson, 1984). Perhaps more consistent with a modulatory PFC, however, is "environmental dependency syndrome". In this symptom profile, merely seeing a stimulus initiates the behavior associated with its use, despite potentially extreme contextual inappropriateness (e.g., eating food from someone's table when walking through a restaurant; D'Esposito, 2003). Furthermore, some monkey lesion studies seem supportive of a modulatory PFC, such as FEF lesions disrupting the ability to mitigate interference from previous trials during oculomotor delayed response (Tsujimoto and Postle, 2012). Similarly, in Rossi et al. (2009), after ablation of lateral PFC, monkeys could still perform visual feature-based attention tasks, but only when the target feature was consistent across trials.

Other considerations, however, suggest that the story is more complicated. First, many participants in neuropsychological studies may have a considerable amount of lateral PFC spared, as individuals with complete lesions may be completely unable to interact with the external world (Odegaard et al., 2017), and thus may frequently be excluded from neuropsychological research. Such cases of total loss of percept-driven behavior are difficult to reconcile with a strictly modulatory view of PFC function. Indeed, at least in the Rossi et al. (2009) study, following the lesion, monkeys required extensive retraining in the contralesional hemifield with substantially simplified tasks prior to assessment. This suggests that at least part of the sensory–response pathway had to be established anew, and perhaps even an entirely different, less control-demanding behavioral strategy was adopted (e.g., win–stay, lose–shift).

Clearly, this topic is difficult to address empirically. Ultimately, what is needed to distinguish a modulatory versus transmissive PFC is a detailed understanding of the large-scale organization of the circuitry that mediates perception–action pathways, and the position of PFC within this circuitry. We discuss this further in the Broader Issues section below.

3 Adaptive Coding through Random Mixed Selectivity

Contemporaneous with development of Guided Activation, another perspective on the function of PFC, Adaptive Coding, was put forth by Duncan (2001), and Duncan and Miller (2001). Based on neurophysiological and neuroimaging findings, Duncan argued that lateral PFC functions as a highly general-purpose problem solver, particularly for difficult tasks, that adapts to the specific demands of the current task through unspecified mechanisms of rapid plasticity (i.e., adaptive coding; Duncan, 2001).

3.1 Adaptive coding perspective

The idea that PFC is a general computational resource, capable of acquiring selectivity to the demands of near any task, was advanced with two lines of evidence. First was human functional

neuroimaging. A meta-analysis of approximately two dozen studies indicated that ostensibly diverse task contrasts — for example, involving the level of response conflict, amount of practice, load in working memory, and perceptual difficulty — all led to peak activations in relatively proximal areas in lateral frontal cortex, as well as within dorsal anterior cingulate cortex (Duncan and Owen, 2000).

The second piece of evidence was from monkey neurophysiology, in that lateral PFC neurons can rapidly acquire selectivity to task-relevant information, yet at the same time, often display difficult-to-characterize *conjunctive coding*. Such conjunctive coding can be understood as a neural response that depends on the interaction between distinct task factors. Take, for example, the rule-based target detection task of White and Wise (1999), an experimental design in which a rule factor (spatial, associative) was crossed with four target locations (left, right, up, down). The neurons that displayed abstract coding of the task rule, which were the focus of the report (reviewed above), were significantly more active when one rule was in play than another, generally across the different cued locations. From a multiple regression or ANOVA framework, these neurons displayed a main effect of rule. A conjunctive coding neuron, however, would display an interaction (perhaps in addition to a main effect), in which the neural response would depend on the non-linear combination of rule and cued location. In other words, a conjunctive coding neuron would respond to a certain cued location (or set of locations), but only (or particularly) while certain rules were in play. In their study, White and Wise (1999) found that many neurons conjunctively coded for rule*location combinations. Furthermore, selectivity of neurons seemed also to depend on the specific epoch of the trial. For example, early in the trial, some cells showed a selectivity to (main effect of) the spatial rule, but late in the trial, showed a preference for the associative rule. Thus, given various task factors (epoch, rule, cued location), many possible profiles of selectivity are possible for a neuron to exhibit. Intriguingly, despite the high number of possibilities, "virtually every conceivable combination was observed" (White and Wise, 1999).

Rather than the exception, such observations of complex conjunctive coding in lateral PFC seem to be the rule. For example, in Wallis et al. (2001), although about half of the recorded neurons were on average selective to one rule (match or non-match), a third were selective to (i.e., conjunctively coded) rule*cue (juice, color) combinations (i.e., fired more to a specific cue for a specific rule). Other early examples of conjunctive coding abound, such as between stimulus features (direction and strength of motion) and action (saccade direction; Kim and Shadlen, 1999), stimulus identity and location (Rao et al., 1997), and actions and outcomes (Wallis and Miller, 2003; cf. Passingham and Wise, 2012).

Combining these two general findings, Duncan (Duncan, 2001; Duncan and Miller, 2001) proposed that lateral PFC uses unspecified mechanisms of rapid plasticity, or adaptive coding, to function as a "global working memory" or "global attentional system", that can be flexibly devoted to whatever task is at hand. In fact, adaptive coding in these areas was proposed to underlie the construct of general intelligence. A long-standing finding from psychometrics research is that individual's performance on cognitive tasks occurs on a "positive manifold", whereby positive correlations tend to be observed across any two task pairings (Spearman, 1904). This highly general factor of individual difference matched the general-purpose nature of Adaptive Coding.

Subsequently, neuroimaging work has replicated the basic findings of Duncan and Owen (2000) with increasingly sophisticated methodology (Assem et al., 2020; Fedorenko et al., 2013; Niendam

et al., 2012; Poldrack, 2011; Woolgar et al., 2016). For example, Fedorenko et al. (2013) administered a battery of various demanding tasks and found that a common set of regions — including caudal lateral PFC, a large patch in rostral DLPFC, lateral posterior parietal cortex, and dorsomedial PFC regions — were activated on difficult versus easy trials, generally within each participant (Fedorenko et al., 2013). Recently, Assem et al. (2020) examined data from a sample of 449 participants in the Human Connectome Project, each of which completed three tasks (an N-back working memory, arithmetic, and a relational reasoning task, which involved finding an abstract rule that correctly classifies two sets of stimuli; Assem et al., 2020). Similar cortical areas were generally activated by difficult conditions across the tasks, but now three separate lateral PFC clusters could be identified at the group-level: ventrocaudal PFC (including IFJ), caudal DLPFC (perhaps 9/46d), and rostral DLPFC (perhaps 46), as well as a more dorsal area, rostral to putative FEF. Fittingly, these regions have been termed the "multiple demand" (MD) network, all of which are all hypothesized to display adaptive coding properties and to function as a cohesive network during novel problem-solving tasks (Duncan, 2010; Duncan et al., 2020).

3.2 Random Mixed Selectivity model: Core hypotheses

Several years later, a neural network model of how such adaptive coding may be implemented, Random Mixed Selectivity, was developed (Rigotti et al., 2010, 2013). This model essentially posited that the representational scheme of lateral PFC was relatively unstructured. Moreover, the Random Mixed Selectivity model illustrated the benefits that such a scheme can have for rapid task learning and flexibility, given a limited set of neurons that must be reused across tasks. Because of the complexity of the model, we next walk through it in detail.

Rigotti et al. (2010) developed their framework within the context of WCST (Figure 3A), using a recurrent neural network (RNN) to model neural representations underlying this task (Figure 3C). The network was recurrent because of the layer of interconnected units with bidirectional (i.e., recurrent) connections. In these networks, the output at one time point is fed back into the network as input at the next time point. As a result, in response to external input, RNNs generate *reverberating* patterns of activity. In certain RNNs, these reverberations can be self-sustaining (i.e., becoming "locked-in" once they emerge), in which case the sustained patterns are termed "attractors". Attractor networks are useful for simulating working memory processes (Wang, 2002) because they can temporarily sustain representations in an active state over extended periods of time. Indeed, Rigotti et al. (2010) conceptualized the sequence of events within each trial to be a series of four active states in working memory. These states are illustrated for an example trial in Figure 3A (and correspond to the black path in Figure 3B). In the RNN, each of these states was encoded by a different fixed-point attractor (i.e., a single, self-sustaining pattern of activity) within the layer of recurrent neurons. Thus, on each trial, the recurrent layer would traverse through a sequence of four attractors. Transitions among attractors were driven by external inputs that signaled color, shape, and outcome information. The goal of training was to let these external inputs drive the recurrent activity along the correct sequence of attractors, so that at the end of the trial, the network reached the attractor corresponding to the correct response. But because the active rule could secretly switch across trials, the network had to learn to use the outcome information to infer the active rule for the next trial. Specifically, when the outcome units signaled an error had been committed, the network had to learn to switch its representation to the opposite rule (i.e., shift the attractor from one rule to another) for the next trial.

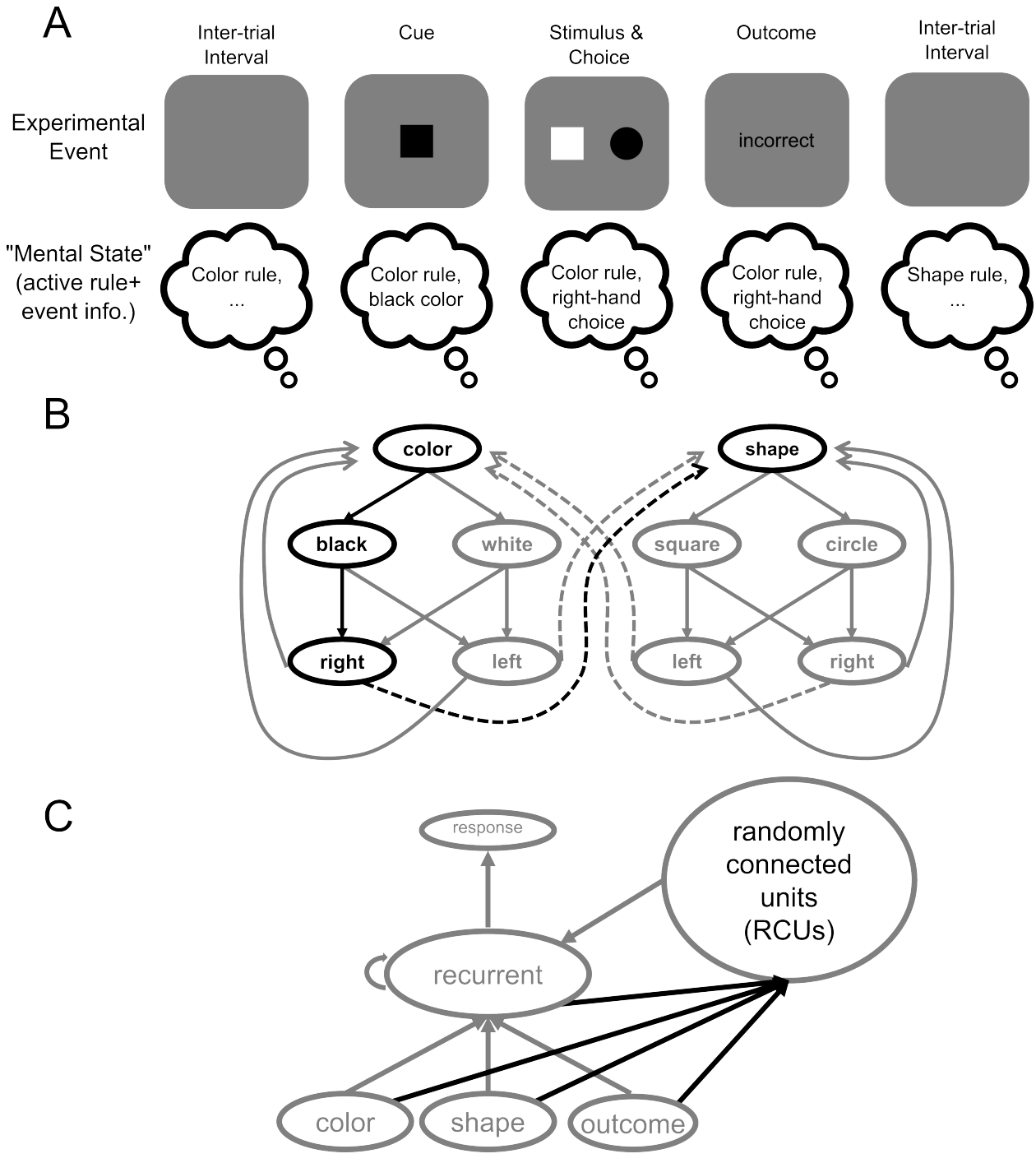

Figure 3: See following page for caption.

Figure 3: **Schematic of the Random Mixed Selectivity model of Wisconsin Card Sorting Task (adapted from Rigotti et al., 2010). A.** Adapted from Rigotti et al. (2010), a trial sequence from the WCST. In this example trial (upper panels), the subject must select a probe stimulus (white square or black circle on third screen) on the basis of a preceding cue stimulus (black square), and a rule, held in mind ("match by color" or "match by shape"). Unbeknown to the subject, the rule may switch (or repeat) from trial to trial. Subjects must use the outcome (i.e., error) signal to infer when the rule has switched, and correspondingly update their task set on the subsequent trial. Rigotti et al. (2010) conceptualized each epoch of trials within this task as corresponding to a discrete state within working memory (lower panels). **B.** Cartoon depiction of all possible transitions among states. Each oval corresponds to a discrete state in working memory (a fixed-point attractor within the model). External events drive transitions among working memory states. Highlighted in black are the transitions that would be implemented by the sequence in A. Dashed lines indicate rule switches — transitions that the network could not implement without an additional layer of units that encode outcome*rule conjunctions. **C.** Schematic of model structure. Black connections —– inputs to the RCUs —– were random and fixed (not modified with training). All grey connections were modified with training. The recurrent and RCU layers together are the task representation. This model is transmissive because the task representation is embedded within the pathway from stimulus to response.

However, as highlighted by Rigotti et al. (2010), structuring the network in this way —– with only input and recurrent units —– leads to a problem in switching among the task rules. To correctly switch rules after an error, the outcome units must drive the recurrent activity towards the attractor of the non-activated rule (dashed lines in Figure 3B). The identity of the non-activated rule, however, differs depending on the current rule: when the shape rule is active, the network must switch to the color rule; when the color rule is active, it must switch to the shape rule. To switch appropriately, therefore, the network must somehow be trained such that the outcome units can push the recurrent activity in *opposite directions* on shape versus color trials. This requirement, however, poses an intractable problem for this network. Because the input that the recurrent units receive from the outcome units is the same in both contexts (i.e., the outcome units do not "know" the current rule), the network cannot learn to respond differently to errors based on the rule.

This problem is well known in computing as the XOR ("exclusive or") problem, which reared its head because the recurrent neurons only received *linear combinations* of outcome and rule variables. To state the problem differently, the recurrent units could not tell apart the error signal during the color rule, from the error signal during the shape rule, because the error signal was the same in both contexts. But, if the error signal was somehow *integrated* with the rule representation before being fed into the recurrent layer, such that two separate error signals could be provided —– one for the color rule, the other for the shape rule —– then the network could actually learn to switch among the rules appropriately. Rigotti et al. (2010) solved this problem by adding a large pool of units to the network, termed "randomly connected units" (RCUs), which provided the recurrent units with the integrated outcome*rule signals they needed for successful switching dynamics (Figure 3C). To do this, these RCUs received projections from the outcome and recurrent units and projected back to the recurrent units.

A useful means of reasoning simply about the XOR issue, and the impact that these RCUs have on the network, is with a geometric perspective and graphical visualization. Geometrically, the

activity of each neuron corresponds to a single dimension. Thus, a pattern of activity across a set of three neurons is a point in a three-dimensional "neural state space" (Figure 4, left; Ebitz and Hayden, 2021). Say one neuron is an *outcome neuron* (e.g., fires when error feedback is delivered), the second is a *rule neuron* (e.g., fires when the shape rule is active), and the third responds to outcomes or to rules, an *outcome+rule* neuron (e.g., fires when the shape rule is active or when error feedback is delivered). At the end of the trial sequence when outcome feedback is received, this three-neuron population will respond with one of four patterns. These four patterns are the points (black and grey spheres) in Figure 4, left.

Importantly, these three neurons encode *linear combinations* of the outcome and rule variables (equivalently, these neurons have *linear selectivity*) —— that is, the expected firing rate of each of these neurons, R, is a weighted sum of the outcome and rule factors[1]:

$$R = b_1 \cdot x(\text{outcome}) + b_2 \cdot x(\text{rule})$$

The terms b_1 and b_2 are weights that indicate how the neuron responds to the respective variable. (Note that for the *rule neuron*, $b_1 = 0$, and for the *outcome* neuron, $b_2 = 0$). In order to consider the impact of this linear encoding of outcome and rule variables, these neurons can be treated as the inputs to any given "downstream" neuron in the recurrent layer. Let us assume that the function of this downstream neuron is to signal that the active rule for the upcoming trial is the "color" rule. If appropriately trained, this downstream neuron will have a set of weights that ensures that it will fire after trials in which (1) the color rule was successfully used, or (2) the shape rule was unsuccessfully used (the black points in Figure 4, left). Graphically, this corresponds to finding a linear decision boundary that separates input patterns into the appropriate "fire" and "not fire" classes (i.e., in Figure 4, left, drawing a line within the grey plane that puts the black points on one side, and the grey points on the other; Duda et al., 2001). Such a line, however, does not exist.

Now consider what happens when the linearly selective *outcome+rule* neuron is exchanged with a non-linearly selective neuron (an RCU of the network). This new neuron encodes a non-linear conjunction of outcome and rule signals (the *outcome*rule* neuron in Figure 4, right): it responds only to error feedback when the shape rule was used. The equation that describes this neuron's firing rate now needs a third term, for the interaction between outcome and rule variables:

$$R = b_1 \cdot x(\text{outcome}) + b_2 \cdot x(\text{rule}) + b_3 \cdot x(\text{outcome}) \cdot x(\text{rule})$$

In Figure 4, right, the possible input patterns now span three dimensions instead of two. This increase in dimensionality enables a linear boundary (now a plane) to be drawn that indeed makes the appropriate classification possible. The purpose of the RCUs can now be understood: to increase the dimensionality of the task representation within the recurrent layer, with the goal of enabling greater diversity in the possible transitions among attractor states.

[1]The terms "x(outcome)" and "x(rule)" specify the level of the outcome (error, correct) and rule (color, switch) factors for a given trial. For the purpose of this example, we assume a contrast coding scheme [e.g., for error trials, $x(\text{outcome}) = 1$, but for correct trials, $x(\text{outcome}) = -1$], and that the mean firing rate of this neuron is equal to zero (so no intercept coefficient is needed).

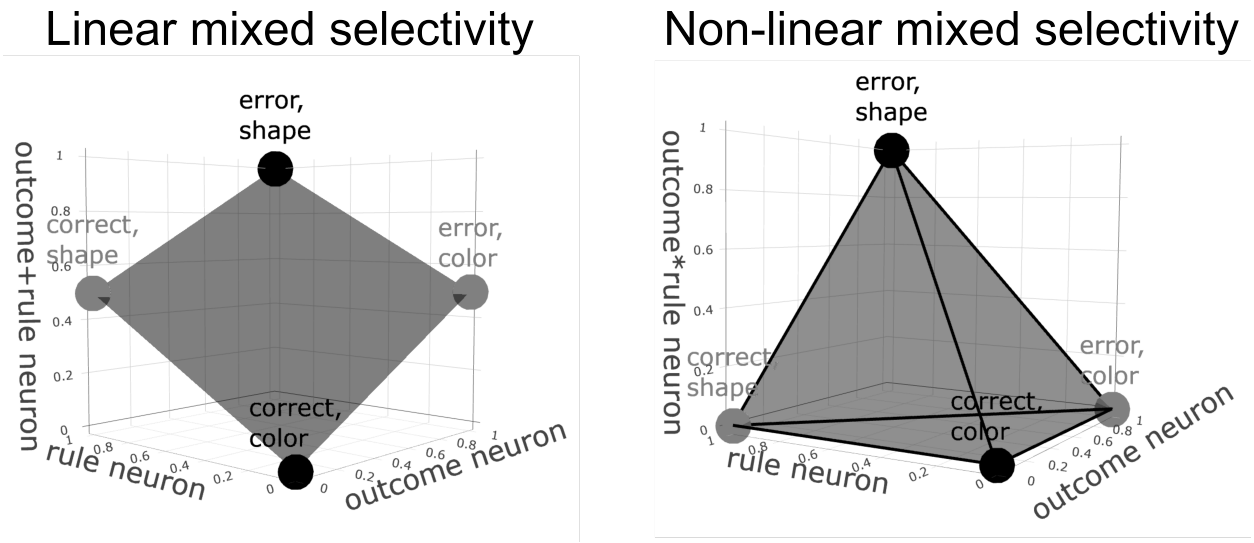

Figure 4: **Cartoon depiction of dimensionality expanding benefits of RCUs (adapted from Fusi et al., 2016). Left (Linear mixed selectivity)**. A three-neuron population consisting of a neuron that only responds to errors (outcome neuron), a neuron that only responds while the shape rule is active (rule neuron), and a neuron that responds to a *linear mixture* of these task features (outcome+rule neuron). Accordingly, the geometry of the patterns span only two dimensions (grey shaded plane). **Right (Non-linear mixed selectivity).** When a linearly selective neuron is replaced with a non-linear mixed selective neuron (outcome*rule), the dimensionality of the geometry expands. Now, any possible classification can be implemented.

The addition of an extra layer to solve an XOR problem was not groundbreaking. What was clever about the solution proposed by Rigotti et al. (2010) was the structure of this additional layer: the additional units received *random and fixed* projections from the input and recurrent units (Figure 3) —– that is, these connections were not modified during training. As a result, these RCUs can be thought of as a large, heterogeneous reservoir of general-purpose "problem solvers". By chance, some units will have pure selectivity to one task feature (e.g., respond only to the shape rule), others will have linear mixed selectivity (e.g., respond to the shape rule *or* to error feedback), and others will have non-linear mixed selectivity (e.g., respond *only if* the shape rule *and* error feedback are present together). These latter units are the critical XOR problem-solvers. Given enough RCUs, there will be enough non-linear mixed selective RCUs to solve all XOR problems demanded by the task. Furthermore, because RCU weights are fixed, these weights do not require training —– that is, any potentially useful conjunction of the RCU and input signals are "already" encoded by the RCUs within an easy-to-read (higher-dimensional) format.

The benefits of random mixed selectivity were demonstrated in a variety of simulations. For instance, increasing the number of RCUs greatly sped up the training process, and also led to less confusable task representations (increasing the distance between attractors in the final model). Importantly, though, as the complexity of the task increased (i.e., more conditions, or more transitions among conditions), the number of required RCUs increased only linearly. Further, across the range of complexity, the number of required RCUs was only on average about three times as great as in a "minimal" task-specific network, which was selectively tuned to the task (i.e., with the RCU weights modified as well). Given that the RCUs were not trained, this efficient scaling behavior is perhaps surprising. One way of reasoning about this efficient scaling is to consider that, of the possible combinations of outcome and rule signals that could be encoded, many such combinations will let the network solve the XOR problem. For example, in the example in Figure 4, right, it actually does not matter if the *outcome*rule* neuron (the RCU) fires in response to *error, shape*, or to *correct, color*. In fact, so long as the RCU is selective to an odd number of conditions (i.e., 1, or 3), it will increase the dimensionality of the representation, permitting an appropriate decision boundary to be learned.[2,3]

[2]Two binary factors, outcome (error, correct) and rule (shape, color), give a contingency table of four cells (i.e., conjunctions of factor levels, or conditions). There are $2^4 = 16$ possible ways to split these conditions into two separate groups (of potentially unequal size). Equivalently, we can consider all possible ways to choose a combination of k conditions from the full set of 4: in binomial coefficient notation, $16 = \Sigma_{k=0}^{4}\binom{4}{k}$. These are the 16 possible combinations of conditions a neuron could hypothetically encode. For example, a neuron could fire in response to (1) no conditions; (2) error*shape; (3) error*shape or error*color, . . . , (16) all conditions. Half of these combinations involve an odd number of conditions: $8 = \binom{4}{1} + \binom{4}{3}$. In a rough sense (assuming all points are equidistant in the geometry), these eight combinations would simply correspond to different rigid-body rotations and reflections of the higher-dimensional configuration of points in Figure 4, none of which would impact the ability of an appropriately wired downstream neuron to decode. Thus, of all the possible profiles of selectivity a neuron could exhibit in this task, 50% would confer the dimensionality-expanding benefits exhibited in Figure 4. Thus, with random connectivity (all profiles of selectivity equally likely), the probability that a neuron would exhibit non-linear mixed selectivity here is $1/2$.

[3]What is a task condition? Rigotti et al. (2010) assume a balanced factorial experimental design, involving the crossing of two or more factors. Thus, the number of task conditions would be equal to the number of cells in the contingency matrix (i.e., product of the number of levels of each factor). But what exactly constitutes a discrete "condition" may not always be clear (e.g., with continuous variables), in which case empirical definitions can be used (Fusi et al., 2016)).

3.3 Key computational mechanisms

In a biological neural population that implements random mixed selectivity, as formalized by Rigotti et al. (2010), three computational properties should be exhibited during performance of a cognitive task. First, the geometry of the neural population should have maximal dimensionality. To understand what this means, consider again the neural state space shown in Figure 4. Each panel shows the patterns of activity across three neurons that were evoked by the task conditions. These patterns of activity form shapes, or geometries, within these mathematical spaces, which define the information that can be decoded from the population response. The number of independent dimensions that the geometry spans (two in the left panel, three in the right) defines the dimensionality. So long as there are more neurons than task conditions, the maximal dimensionality that a geometry can exhibit is bounded by the number of distinct task conditions. To state this another way, under maximal dimensionality, each task condition should evoke a pattern of activity that is not only consistent across trials, but also separable from all other condition-evoked patterns along its own, unique dimension. In order for maximal dimensionality to emerge, two conditions must be met. First, the population must consist of neurons that are selective to (non-linear) conjunctions of different task factors. Second, these selectivities to the conjunctions must be sufficiently heterogeneous so as to "cover all bases". In particular, even conjunctions that are putatively task-irrelevant should be encoded by the population.

The second and third predictions of Random Mixed Selectivity are termed the "pre-eminence" and "universality" features of the model (Rigotti et al., 2010). Evidence of random mixed selectivity should be pre-eminent, in that the neurons should exhibit conjunctive coding of task features even *prior to learning* that the features are task relevant. It should also be universal in the sense that there should always be some neurons selective to the various conditions of any arbitrary task (Rigotti et al., 2010). These features specifically emphasize the intended computational goal of the Random Mixed Selectivity model: to confer flexibility, for example, in rapidly acquiring novel task sets, or in quickly learning to exploit an arbitrary combination of conditions within an existing task set.

3.4 Empirical support

Given the putatively stricter requirements of the Random Mixed Selectivity account, it is important to determine whether the extant data are consistent. Indeed, the random mixed selectivity model can account for aspects of monkey neurophysiology data from DLPFC quite well. As stated above, an often-reported feature of this region is its conjunctive, heterogeneous coding of various task conditions (Assad et al., 1998; Passingham and Wise, 2012; White and Wise, 1999). Yet a key question for the account is whether the conjunctive coding is actually heterogeneous enough across the population to support a high-dimensional code. While the prior data were roughly consistent (White and Wise, 1999), Rigotti et al. (2013) conducted a tour-de-force reanalysis of an existing dataset (Warden and Miller, 2010) to provide the first rigorous test of the account. Monkeys were trained to perform an object sequence memory task, in which the monkey had to remember a target sequence of two objects and report the sequence either via "recognition" (a delayed match-to sample) or "recall" (a saccade between targets within an array). There were 12 possible unique object sequences, each of which appeared in both task types (recognition, recall), forming 24 distinct conditions. Responses from DLPFC (area 46) indeed showed conjunctive coding. For example, many neurons reliably responded to specific stimuli, but only when they appeared

in a specific position of the sequence. To estimate dimensionality, a procedure was developed closely aligned to the theoretical framework. As high-dimensional representations permit many different conjunctions of task variables to be decoded, dimensionality can be estimated by testing all possible binary classifications, then tallying the proportion that could be successfully classified (with 24 conditions there are 224 possible classifications). This brute-force approach was done in a sophisticated manner, using oversampling of neurons and theoretical results, to obtain a robust, unbiased estimate of dimensionality.

In line with predictions of random mixed selectivity, dimensionality was near-maximal after the second cue was presented — that is, each of the 24 conditions effectively evoked an activity pattern that was orthogonal to all other conditions. Additionally, when considering the recognition task alone (in which monkeys made many more errors than the recall task), dimensionality and behavior were tightly linked, such that when the monkey made an error, dimensionality collapsed. But, despite this collapse, the identity of the two cues could still be decoded on error trials with near-perfect accuracy.

Rigotti et al. (2013) provided a striking demonstration of the elegance of Random Mixed Selectivity in accounting for hard-to-interpret properties of single-neuron responses in DLPFC. Errors on this task were strongly and specifically related to a conjunctive format, in which specific stimuli were bound together in a unique code for their sequence. Further, these sequences were encoded differently across task types (recall, retrieval) — that is, the same sequence of images was encoded with a distinct pattern of activity, depending on the task type. In other words, stimulus sequence and task type were mixed non-linearly. Prior to Random Mixed Selectivity, such a finding may have been quite challenging to interpret within a meaningful theoretical framework.

Nevertheless, the results leave room for alternative accounts. Many of the key predictions of Random Mixed Selectivity went untested, such as whether the conjunctive sequence*task coding — perhaps the most surprising finding — emerged obligatorily and prior to training. But perhaps most critically, it is still not known whether the sequence*task-type coding served any apparent functional purpose, as the brain–behavior analysis was conducted only on the recognition task. The collapse of dimensionality therefore reflected only an "unbinding" of the stimulus identities (a loss of sequence information within a single task). With this caveat in mind, the striking nature of this result may be somewhat tempered, given the established importance that object sequences seem to play in DLFPC coding (Genovesio et al., 2009, 2011; Passingham and Wise, 2012). In other words, it is not apparent that the Random Mixed Selectivity framework is currently required to explain the brain–behavior correlations that are related to the multi-task (task-switching) nature of the paradigm.

4 Broader Issues

Two influential models of neural task set representations have been discussed: Guided Activation and Adaptive Coding. The conceptual framework for each of these models was developed contemporaneously in the early 21st century, based on overlapping bodies of cognitive neuroscience research. Thus, they have many similarities. Nevertheless, subsequent theoretical work (Rigotti et al., 2010; Rougier et al., 2005) has elaborated upon these frameworks, helping to differentiate them along several important dimensions. Here these differences are summarized, in terms of

their broader implications for theories of task representation, and in some cases, with suggestions as to how the two perspectives might be reconciled.

4.1 Abstract versus high-dimensional representations

The task representations of Adaptive Coding (Rigotti et al., 2010) differ from the representations of Guided Activation (at least those that have been specified, e.g., Cohen et al., 1990; Rougier et al., 2005) in their level of abstraction. In the Guided Activation account, task representations are highly abstract, representing task-relevant rules or stimulus dimensions that *generalize* across different aspects of the task (e.g., task-irrelevant features). This abstraction also makes them relatively low-dimensional. In contrast, in the Adaptive Coding account, task representations are highly conjunctive and heterogeneous (across the population), representing many different non-linear combinations of rules, stimuli, and other task aspects that, collectively, *discriminate* among all the different conjunctions of the task, including task-irrelevant features. These features make the code relatively high-dimensional.

Human behavior exploits abstractions. In particular, people spontaneously generate abstract sets to guide cognitive processing (Collins and Frank, 2013), and have a strong bias towards maintaining previously used sets, even when instructed to do otherwise (Dreisbach, 2012). But at least in monkey neuronal data, the extant findings suggest both that abstractions and conjunctions are represented in DLPFC. How can these clearly antagonistic representational formats be reconciled?

DLPFC could use a combination of both abstract and conjunctive representations. One variable that may distinguish the use of abstract versus high-dimensional (conjunctive) codes is the stage of learning regarding the relevant task-sets (Badre et al., 2021). The benefits of an abstract format are straightforward to consider in novel settings —– early during task-set acquisition, when the participant has limited, or even no, prior direct experience with the specific task conditions. In these settings, the participant can draw upon past experience —– that is, from other, similar tasks — to infer how to perform the current one. Similarity-based inference would depend on abstract, lower-dimensional representations. Use of such processes and representations could account for the enormous flexibility with which humans can perform novel tasks, such as those that require quick learning or the use of arbitrary, context-dependent associations between task elements (Cole et al., 2011, 2013).

In contrast, the benefits of a high-dimensional task representation may emerge when the basic components of the task set are already well-learned and proceduralized, or when flexible control processes are demanded, such as switching among different aspects of the task. When dimensionality increases, similarities rapidly decrease (this is one symptom of the "curse of dimensionality"; Duda et al., 2001), which would impare generalizability and inference. But, as reviewed earlier, representing the task set in higher dimensions could facilitate flexible switching behavior, by increasing discriminability of distinct task contexts (Fusi et al., 2016; Rigotti et al., 2010).

The notion that higher-dimensional coding may emerge (or at least become particularly beneficial) once a task set is already established, may also help in defining the boundary conditions on conjunctive coding and dimensionality. Although Random Mixed Selectivity does emphasize "pre-eminence" and "universality", unless an unrealistically large representational capacity is to be invoked, there must be some constraint on the different variables that are conjunctively encoded with one another. One such constraint may depend on practice with the task. Early

in practice, abstract goal-dependent representations (e.g., rules or task contexts could play a larger role). Conjunctive coding might then develop with respect to these repeatedly reactivated representations. For example, salient external events that occur while a certain task context is active could be encoded as an event*context conjunction, which could become increasingly emphasized in the code through incremental learning mechanisms (e.g., Hebbian or reinforcement learning). Indeed, this general view that the preponderance of non-linear mixed selectivity in DLPFC is enhanced with experience with the task is in line with recent theoretical (Lindsay et al., 2017) and preliminary empirical work (Dang et al., 2021), that demonstrate that the proportion of non-linearly selective neurons is increased, at least in some scenarios, by practice.

4.2 Modulatory versus transmissive PFC

The two accounts of task representation also differ in their view of PFC as a modulator or a transmitter (Badre et al., 2021). The Guided Activation account assumes that PFC modulates distal representations which "actually" perform the task (Miller and Cohen, 2001), while the Random Mixed Selectivity account assumes that PFC integrates its input with internal state information and directly transmits the selected action to downstream premotor circuitry (Rigotti et al., 2010). These alternative views concern the causal path that the flow of information takes during controlled behavior. Resolving this debate is therefore a complex issue that will ultimately demand a cohesive computational theory built on detailed and comprehensive knowledge of prefrontal connectivity. Though cohesive theories based on what is known about connectivity certainly exist (Miller and Cohen, 2001; Passingham and Wise, 2012), the requisite detail of this anatomical knowledge for resolving this debate is currently lacking.

Perhaps because of this complexity, modulatory–transmissive debates seem to be an extensive divide in the cognitive neuroscience of PFC. Various areas of frontal cortex have been proposed, by different investigators, to serve either a modulatory or transmissive role. For instance, an early (and yet ongoing) debate over DLPFC concerns its role in sensory working memory: does it serve to maintain the sensory content of working memory (Constantinidis et al., 2001; Courtney et al., 1997; Goldman-Rakic, 2011; Xu, 2017) once it can be *transmitted* to motor-related structures, or does it serve to mitigate interference by *modulating* "sensory stores" elsewhere (D'Esposito and Postle, 2015; Petrides, 2000; Postle, 2006)? Perceptual decision-making models of DLPFC also have a transmissive flavor, in that they propose the role of DLPFC and associated "decision circuitry" (involving FEF, LIP, superior colliculus, and other areas) is to accumulate sensory information and transform it into a decision (Kim and Shadlen, 1999). In the medial wall, over dorsal anterior cingulate cortex (dACC), a similar divide of perspectives is present (Ebitz and Hayden, 2016; Heilbronner and Hayden, 2016). Here, modulatory models propose that dACC functions as an auxiliary monitor or regulator, which modulates control-related coding in other areas, such as DLPFC (Botvinick et al., 2001; Shenhav et al., 2013). Conversely, transmissive accounts of dACC function cast the region as one integrated component of distributed decision-making pathways that gradually transform choice options into actions (Heilbronner and Hayden, 2016; Kolling et al., 2016).

As indicated above, the lesion and stimulation evidence with respect to lateral PFC paints a mixed story. Some reports of focal lesions in lateral PFC affect behavioral errors consistent with a modulatory role (Petrides, 2000; Tsujimoto and Postle, 2012), in that the ability to overcome interference is impacted, yet there is not necessarily disruption to all components of the task set.

Conversely, the data regarding more widespread PFC lesions suggest a disruption of behavior that is more consistent with that predicted by transmissive roles (Odegaard et al., 2017). Similar ambiguity is associated with stimulation experiments, in that some studies have indeed indicated that stimulating lateral PFC has causal impacts on coding in sensory regions, and on behavioral measures of attention (Moore and Armstrong, 2003). Yet, this stimulation could have impacted both internal coding in PFC, as well as downstream response-related coding. Thus, we do not know if the observed sensory enhancement was causal to the behavioral change, which would be predicted from a modulatory view. While these issues illustrate the potential complexity of adjudicating transmissive versus modulatory architectures with lesion and stimulation methods, it is possible that careful combinations of both methods (perhaps with multi-site recording, whole-brain neuroimaging, or tract disconnection) could afford stronger inference.

One challenge to modulatory models is the parsimony that can be achieved with transmissive models, which do not have to invoke a dedicated, external controller that is separate from the input–output pathway (see Ehrlich and Murray, 2021, for a similar argument). The Random Mixed Selectivity account demonstrates one example of a transmissive framework. Another is provided by a study from Mante et al. (2013), which also employed an RNN, trained to perform a cued task-switching task (switching among classifying the direction moving dots, or their color). Their model was transmissive in that sensory input was delivered to the recurrent layer, and responses were also read out of it. In the trained model, control over response selection (switching) was enforced by the same circuitry that accumulated perceptual evidence towards a decision. Interestingly, the model representations match properties of DLPFC and FEF representations, such as heterogeneity of neural selectivity, from monkeys performing the same task.

While transmissive versus modulatory views are clearly pervasive, the topic is complex and has not yet been a question of direct investigation. Factors such as the nature of the task could influence the appropriateness of modulatory versus transmissive views. If the stimulus–response mappings of a task are arbitrary and less practiced, for instance, then control may play more of a transmissive role. It is also important to note that the dimensionality of a representation is not necessarily diagnostic for transmissive versus modulatory architectures. Though modulatory task representations have been formalized as abstract, they could also be high dimensional (e.g., a high-dimensional modulatory task representation could bias response representations differently depending on the particular conjunction of target and distractor stimulus features). One potentially promising direction for future research would be to explicitly pit them against each other, through RNN models with architectures that are either modulatory or transmissive. Such a study may better illustrate the kinds of representations that would be predicted to emerge under each scheme. This could also afford comparison in the ability of the models to capture aspects of behavioral and neural data.

4.3 Topographic organization of lateral PFC and stability of representations

Another dimension that could potentially differentiate these perspectives is in determining whether lateral PFC has a coherent topographic functional organization. Generally, topographic organization is considered to emerge due to wiring constraints. Long-range versus local connectivity is less space efficient, running into substantial scaling problems with realistic levels of connectivity, and may be less robust for the development of circuits implementing lateral inhibition (Kaas, 1997; Knudsen et al., 1987; Purves et al., 1992). Under such constraints, if certain patterns of input con-

nections are overrepresented in the neural population, one would expect the neurons that receive these connections to be spatially clustered within cortical regions.

The Adaptive Coding account suggests the lack of coherent topography in lateral PFC, a stance which is also emphasized by the Random Mixed Selectivity model. As the input connections to the RCUs in the model are random, all patterns of input connections are equally likely, which would suggest no need for topographic organization. Some transmissive perspectives discussed earlier have similar suggestions, by postulating smooth gradients of functional organization rather than sharp boundaries. Conversely, models positing abstract coding of control that were developed within the framework of Guided Activation suggest a more coherent topography, as connectivity in these models is decidedly non-random, and actually quite interpretable (Rougier et al., 2005).

Another way of conceptualizing these two accounts with respect to connectivity and topography is provided by the distinction recently made between "Sherringtonian" and "Hopfieldian" views of neural systems (Barack and Krakauer, 2021). After his work on simple reflex arcs, the Sherringtonian view posits that the "first-level explainers" (the elements considered to cause a given phenomenon) are individual neurons or units in a network with detailed, weighted connections between them (Sherrington, 1906). Explanations are sought in terms of computations performed by these neurons or units and their patterns of connectivity. The Sherringtonian view has much in common with formulations of the Guided Activation account. Contrastingly, after Hopfield's (1982) work on recurrent neural networks, the Hopfieldian view posits that the first-level explainers are instead properties of neural state spaces (i.e., the network's collective activity). Here, the detailed connectivity patterns of units in the network are of secondary importance to the movement within or between neural state spaces — that is, explanations are not sought in terms of the patterns of connectivity, as these might reflect a level of analysis that is too granular from which to find meaning. Indeed, Random Mixed Selectivity has been highlighted as a prime example of the Hopfieldian view (Barack and Krakauer, 2021).

But even despite these differences, both the Adaptive Coding and Guided Activation perspectives emphasize a prime role of plasticity in lateral PFC organization (Duncan, 2001; Miller and Cohen, 2001). In either scheme, lateral PFC must receive and send information to other heteromodal cortices that are themselves subject to plasticity (Miller and Cohen, 2001). To handle this, plasticity mechanisms could involve modification or development of new synapses or of new neurons (Gould et al., 1999; Miller and Cohen, 2001). Thus, even if a coherent topography can be mapped, the longer-term stability of such an organization is another important question.

Yet, in spite of the stance that DLPFC (or PFC more generally) eludes description by a single, unifying function or by the type of information it encodes (Duncan, 2001), the evidence reviewed here suggests that at least one type of information is particularly important to DLPFC: goal-relevant order or sequence information during episodes of interacting with the environment (Genovesio et al., 2009, 2011; Passingham and Wise, 2012; Rigotti et al., 2013). Often in behavior, an event (e.g., stimulus, action, outcome) will define what goals or action contingencies are available in the future. Under view that DLPFC functions to generate and segregate contexts for organizing subsequent behavior (Badre and Nee, 2018; Fuster, 2001), such contexts may naturally be afforded by sequence information, which may be capitalized by DLPFC. This consideration could explain the apparently obligatory encoding of event order observed in DLPFC neuronal activity (Genovesio et al., 2009). In other words, though this view is quite incomplete, it seems reasonable to hypothesize that DLPFC

is indeed "tuned" to a particular kind of information, which is often naturalistically reflected in order or sequence information (e.g., see Ehrlich and Murray, 2021, for one such representational scheme). Thus, it is an open question whether or not order, sequence, or timing-related information —– or whatever information primates have tended to exploit to generate behavioral contexts —– is coherently mapped to single neurons or reflected in the stable topographic organization of DLPFC. However, there does not yet seem to be a compelling reason to abandon attempts to formulate coherent hypotheses regarding these questions, nor to posit that this part of the brain eludes a unified functional description.

Indeed, an intriguing aspect of Guided Activation perspectives that propose abstract rule representations underlie controlled behavior, is that they suggest the potential for discovering, within the functional organization and representational structure of prefrontal cortex, the "vocabulary" of tasks we have typically performed over our evolutionary history (Botvinick et al., 2015). In practice, of course, this question is quite challenging, because of the nearly infinite ways in which tasks can vary. Thus, it remains exceedingly difficult for researchers to decide exactly which dimensions are most profitable to manipulate within experiments. Without strong hypotheses to constrain experimental designs, it may be premature to expect to find coherent task dimensions within the organization of prefrontal representations.

Ideally, these hypotheses would be based on a well-developed, quantitative understanding of how naturalistic planned behaviors tend to vary. Indeed, approaches that have found great success in visual neuroscience emphasize the importance of understanding how the input signal naturalistically varies (Botvinick et al., 2015; Field, 1987). In vision, these input signals are well-studied and quantitative models have been used to learn latent dimensions of naturalistic images; these provide a powerful basis for examining neural response profiles and topographic organization. With this general approach, great strides have been made in understanding not only the representational and topographic organization of visual cortex, including both striate (Gallant et al., 1993; Olshausen and Field, 1996) and extrastriate areas (Gomez et al., 2019), including even higher-order visual areas, such as IT cortex (Bao et al., 2020). Of course, the endeavor is much more complex when considering higher-cognitive regions such as DLPFC, where the "input" signal is highly complex due to strong recurrence. Nevertheless, data-driven approaches such as analyzing neural network representations (Rougier et al., 2005; Yang et al., 2019a, b), or naturalistic ways to mine the "output" —– that is, behavior (e.g., Hayden et al., 2021; Stout et al., 2018; Voloh et al., 2021) —– may provide critical analytic tools for bootstrapping a working model of task spaces, and consequently should be further explored.

5 Summary and Conclusions

In the pursuit of understanding cognitive control and how the human brain achieves it, cognitive neuroscientists have turned to computational models. These models generally embody the assumption that control is enabled by *internal task representations* encoded in the prefrontal cortex, which organize the necessary information for successful task performance. Currently, two influential models of task representations predominate research in this area. Early models of task representations (Cohen et al., 1990) have evolved into the comprehensive theory of Guided Activation (Miller and Cohen, 2001), which emphasizes the role of lateral prefrontal cortex to acquire and use abstract task representations (Rougier et al., 2005). Conversely, more recent models emphasize

advantages of leveraging high-dimensional representational spaces for cognitive flexibility (Rigotti et al., 2010, 2013). These divergent perspectives highlight novel issues of how the human brain represents tasks, which may stimulate progress towards understanding the core mechanisms of cognitive control.

References

[1] Alexander, W. H., & Brown, J. W. (2011). Medial prefrontal cortex as an action-outcome predictor. *Nature Neuroscience*, 14(10), 1338–1344.

[2] Asaad, W. F., Rainer, G., & Miller, E. K. (1998). Neural Activity in the Primate Prefrontal Cortex during Associative Learning. *Neuron*, 21(6), 1399–1407. https://doi.org/10.1016/S0896-6273(00)80658-3

[3] Assem, M., Glasser, M. F., Van Essen, D. C., & Duncan, J. (2020). A domain-general cognitive core defined in multimodally parcellated human cortex. *Cerebral Cortex*, 30(8), 4361–4380.

[4] Badre, D. (2008). Cognitive control, hierarchy, and the rostro–caudal organization of the frontal lobes. *Trends in Cognitive Sciences*, 12(5), 193–200.

[5] Badre, D., Bhandari, A., Keglovits, H., & Kikumoto, A. (2021). The dimensionality of neural representations for control. *Current Opinion in Behavioral Sciences*, 38, 20–28.

[6] Badre, D., & Nee, D. E. (2018). Frontal cortex and the hierarchical control of behavior. *Trends in Cognitive Sciences*, 22(2), 170–188.

[7] Bao, P., She, L., McGill, M., & Tsao, D. Y. (2020). A map of object space in primate inferotemporal cortex. *Nature*, 583, 103–108.

[8] Barack, D. L., & Krakauer, J. W. (2021). Two views on the cognitive brain. *Nature Reviews Neuroscience*, 22(6), 359–371.

[9] Botvinick, M. M., Braver, T. S., Barch, D. M., Carter, C. S., & Cohen, J. D. (2001). Conflict monitoring and cognitive control. *Psychological Review*, 108(3), 624–652.

[10] Botvinick, M. M., & Cohen, J. D. (2014). The computational and neural basis of cognitive control: charted territory and new frontiers. *Cognitive Science*, 38(6), 1249–1285.

[11] Botvinick, M. M., Weinstein, A., Solway, A., & Barto, A. (2015). Reinforcement learning, efficient coding, and the statistics of natural tasks. *Current Opinion in Behavioral Sciences*, 5, 71–77.

[12] Braver, T. S., & Cohen, J. D. (2000). On the control of control: The role of dopamine in regulating prefrontal function and working memory. *Attention and Performance*, 18, 712–737.

[13] Brown, J. W., & Braver, T. S. (2005). Learned predictions of error likelihood in the anterior cingulate cortex. *Science*, 307, 1118–1121.

[14] Bunge, S. A., Kahn, I., Wallis, J. D., Miller, E. K., & Wagner, A. D. (2003). Neural circuits subserving the retrieval and maintenance of abstract rules. *Journal of Neurophysiology*, 90(5), 3419–3428.

[15] Buschman, T. J., & Miller, E. K. (2007). Top-down versus bottom-up control of attention in the prefrontal and posterior parietal cortices. *Science*, 315, 1860–1862.

[16] Chen, N.-H., White, I., & Wise, S. (2001). Neuronal activity in dorsomedial frontal cortex and prefrontal cortex reflecting irrelevant stimulus dimensions. *Experimental Brain Research*, 139(1), 116–119.

[17] Cohen, J. D., Dunbar, K., & McClelland, J. L. (1990). On the control of automatic processes: A parallel distributed processing account of the Stroop effect. *Psychological Review*, 97(3), 332–361. https://doi.org/10/cfn9t7

[18] Cole, M. W., Etzel, J. A., Zacks, J. M., Schneider, W., & Braver, T. S. (2011). Rapid transfer of abstract rules to novel contexts in human lateral prefrontal cortex. *Frontiers in Human Neuroscience*, 5. doi: 10.3389/fnhum.2011.00142

[19] Cole, M. W., Laurent, P., & Stocco, A. (2013). Rapid instructed task learning: A new window into the human brain's unique capacity for flexible cognitive control. *Cognitive, Affective, & Behavioral Neuroscience*, 13(1). doi: 10.3758/s13415-012-0125-7

[20] Collins, A. G. E., & Frank, M. J. (2013). Cognitive control over learning: Creating, clustering and generalizing task-set structure. *Psychological Review*, 120(1), 190–229.

[21] Constantinidis, C., Franowicz, M. N., & Goldman-Rakic, P. S. (2001). The sensory nature of mnemonic representation in the primate prefrontal cortex. *Nature Neuroscience*, 4(3), 311–316.

[22] Courtney, S. M., Ungerleider, L. G., Keil, K., & Haxby, J. V. (1997). Transient and sustained activity in a distributed neural system for human working memory. *Nature*, 386, 608–611.

[23] Dang, W., Jaffe, R. J., Qi, X.-L., & Constantinidis, C. (2021). Emergence of nonlinear mixed selectivity in prefrontal cortex after training. *Journal of Neuroscience*, 41(35), 7420–7434.

[24] deCharms, R. C., & Zador, A. (2000). Neural representation and the cortical code. *Annual Review of Neuroscience*, 23(1), 613–647.

[25] Desimone, R., & Duncan, J. (1995). Neural mechanisms of selective visual attention. *Annual Review of Neuroscience*, 18(1), 193–222.

[26] D'Esposito, M. (2003). (Ed.). *Neurological foundations of cognitive neuroscience*. Cambridge, MA: MIT Press.

[27] D'Esposito, M., & Postle, B. R. (2015). The cognitive neuroscience of working memory. *Annual Review of Psychology*, 66(1), 115–142.

[28] Dreisbach, G. (2012). Mechanisms of cognitive control: The functional role of task rules. *Current Directions in Psychological Science*, 21(4), 227–231.

[29] Duda, C. R. O., Hart, P. E., & Stork, D. G. (2001). *Pattern classification*. New York: Wiley.

[30] Duncan, J. (2001). An adaptive coding model of neural function in prefrontal cortex. *Nature Reviews Neuroscience*, 2(11), 820–829.

[31] Duncan, J. (2010). The multiple-demand (MD) system of the primate brain: Mental programs for intelligent behaviour. *Trends in Cognitive Sciences*, 14(4), 172–179.

[32] Duncan, J., Assem, M., & Shashidhara, S. (2020). Integrated intelligence from distributed brain activity. *Trends in Cognitive Sciences*, 24(10), 838–852.

[33] Duncan, J., & Miller, E. K. (2001). Cognitive focus through adaptive neural coding in the primate prefrontal cortex. In *Principles of frontal lobe function*. Oxford, UK: Oxford University Press.

[34] Duncan, J., & Owen, A. M. (2000). Common regions of the human frontal lobe recruited by diverse cognitive demands. *Trends in Neurosciences*, 23(10), 475–483.

[35] Ebitz, R. B., & Hayden, B. Y. (2016). Dorsal anterior cingulate: A Rorschach test for cognitive neuroscience. *Nature Neuroscience*, 19(10), 1278–1279.

[36] Ebitz, R. B., & Hayden, B. Y. (2021). The population doctrine in cognitive neuroscience. *Neuron*, 109(19), 3055–3068.

[37] Egner, T., & Hirsch, J. (2005). Cognitive control mechanisms resolve conflict through cortical amplification of task-relevant information. *Nature Neuroscience*, 8(12), 1784–1790.

[38] Ehrlich, D. B., & Murray, J. D. (2021). Geometry of neural computation unifies working memory and planning. *BioRxiv*. doi: 10.1101/2021.02.01.429156

[39] Etzel, J. A., Cole, M. W., Zacks, J. M., Kay, K. N., & Braver, T. S. (2016). Reward motivation enhances task coding in frontoparietal cortex. *Cerebral Cortex*, 26(4), 1647–1659.

[40] Fedorenko, E., Duncan, J., & Kanwisher, N. (2013). Broad domain generality in focal regions of frontal and parietal cortex. *Proceedings of the National Academy of Sciences*, 110(41), 16616–16621.

[41] Field, D. J. (1987). Relations between the statistics of natural images and the response properties of cortical cells. *Journal of the Optical Society of America A*, 4(12), 2379–2394.

[42] Freedman, D. J., Riesenhuber, M., Poggio, T., & Miller, E. K. (2001). Categorical representation of visual stimuli in the primate prefrontal cortex. *Science*, 291, 312–316.

[43] Freund, M. C., Bugg, J. M., & Braver, T. S. (2021). A representational similarity analysis of cognitive control during color-word Stroop. *Journal of Neuroscience*, 41(35) 7388–7402.

[44] Fusi, S., Miller, E. K., & Rigotti, M. (2016). Why neurons mix: High dimensionality for higher cognition. *Current Opinion in Neurobiology*, 37, 66–74.

[45] Fuster, J. M. (2001). The prefrontal cortex–An update: Time is of the essence. *Neuron*, 30(2), 319–333.

[46] Gallant, J. L., Braun, J., & Van Essen, D. C. (1993). Selectivity for polar, hyperbolic, and cartesian gratings in macaque visual cortex. *Science*, 259, 100–103.

[47] Genovesio, A., Tsujimoto, S., & Wise, S. P. (2009). Feature- and order-based timing representations in the frontal cortex. *Neuron*, 63(2), 254–266.

[48] Genovesio, A., Tsujimoto, S., & Wise, S. P. (2011). Prefrontal cortex activity during the discrimination of relative distance. *Journal of Neuroscience*, 31(11), 3968–3980.

[49] Goldman-Rakic, P. S. (1987). Circuitry of primate prefrontal cortex and regulation of behavior by representational memory. In V. B. Mountcastle, R. Plum & S. R. Geiger (Eds.), *Handbook of physiology, Section 1: The nervous system. Vol. 5: Higher functions of the brain* (pp. 373–417). Bethesda, MD: American Physiological Society.

[50] Gomez, J., Barnett, M., & Grill-Spector, K. (2019). Extensive childhood experience with Pokémon suggests eccentricity drives organization of visual cortex. *Nature Human Behaviour*, 3(6), 611–624.

[51] Gould, E., Reeves, A. J., Graziano, M. S. A., & Gross, C. G. (1999). Neurogenesis in the neocortex of adult primates. *Science*, 286, 548–552.

[52] Hayden, B. Y., Park, H. S., & Zimmermann, J. (2021). Automated pose estimation in primates. *American Journal of Primatology*. doi: 10.1002/ajp.23348

[53] Haynes, J.-D., Sakai, K., Rees, G., Gilbert, S., Frith, C., & Passingham, R. E. (2007). Reading hidden intentions in the human brain. *Current Biology*, 17(4), 323–328.

[54] Heilbronner, S. R., & Hayden, B. Y. (2016). Dorsal anterior cingulate cortex: A bottom-up view. *Annual Review of Neuroscience*, 39(1), 149–170.

[55] Hopfield, J. J. (1982). Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the National Academy of Sciences*, 79(8), 2554–2558.

[56] Kaas, J. H. (1997). Topographic maps are fundamental to sensory processing. *Brain Research Bulletin*, 44(2), 107–112.

[57] Kim, J.-N., & Shadlen, M. N. (1999). Neural correlates of a decision in the dorsolateral prefrontal cortex of the macaque. *Nature Neuroscience*, 2(2), 176–185.

[58] Knudsen, E. I., Lac, S., & Esterly, S. D. (1987). Computational maps in the brain. *Annual Review of Neuroscience*, 10(1), 41–65.

[59] Koechlin, E., Ody, C., & Kouneiher, F. (2003). The architecture of cognitive control in the human prefrontal cortex. *Science*, 302, 1181–1185.

[60] Kolling, N., Wittmann, M. K., Behrens, T. E. J., Boorman, E. D., Mars, R. B., & Rushworth, M. F. S. (2016). Value, search, persistence and model updating in anterior cingulate cortex. *Nature Neuroscience*, 19(10), 1280–1285.

[61] Lindsay, G. W., Rigotti, M., Warden, M. R., Miller, E. K., & Fusi, S. (2017). Hebbian learning in a random network captures selectivity properties of the prefrontal cortex. *Journal of Neuroscience*, 37(45), 11021–11036.

[62] MacDonald, A. W., Cohen, J. D., Stenger, V. A., & Carter, C. S. (2000). Dissociating the role of the dorsolateral prefrontal and anterior cingulate cortex in cognitive control. *Science*, 288, 1835–1838.

[63] Mante, V., Sussillo, D., Shenoy, K. V., & Newsome, W. T. (2013). Context-dependent computation by recurrent dynamics in prefrontal cortex. *Nature*, 503, 78–84.

[64] Maunsell, J. H. R. (2015). Neuronal mechanisms of visual attention. *Annual Review of Vision Science*, 1(1), 373–391.

[65] Maunsell, J. H. R., & Treue, S. (2006). Feature-based attention in visual cortex. *Trends in Neurosciences*, 29(6), 317–322.

[66] Miller, E. K., & Cohen, J. D. (2001). An integrative theory of prefrontal cortex function. *Annual Review of Neuroscience*, 24(1), 167–202.

[67] Miller, E. K., Nieder, A., Freedman, D. J., & Wallis, J. D. (2003). Neural correlates of categories and concepts. *Current Opinion in Neurobiology*, 13(2), 198–203.

[68] Monsell, S. (2003). Task switching. *Trends in Cognitive Sciences*, 7(3), 134–140.

[69] Monsell, S. (2017). Task set regulation. In T. Egner (Ed.), *The Wiley handbook of cognitive control* (pp. 29–49). New York: Wiley/Blackwell.

[70] Moore, T., & Armstrong, K. M. (2003). Selective gating of visual signals by microstimulation of frontal cortex. *Nature*, 421, 370–373.

[71] Morishima, Y., Akaishi, R., Yamada, Y., Okuda, J., Toma, K., & Sakai, K. (2009). Task-specific signal transmission from prefrontal cortex in visual selective attention. *Nature Neuroscience*, 12(1), 85–91.

[72] Muhammad, R., Wallis, J. D., & Miller, E. K. (2006). A comparison of abstract rules in the prefrontal cortex, premotor cortex, inferior temporal cortex, and striatum. *Journal of Cognitive Neuroscience*, 18(6), 974–989.

[73] Nieder, A., Freedman, D. J., & Miller, E. K. (2002). Representation of the quantity of visual items in the primate prefrontal cortex. *Science*, 297, 1708–1711.

[74] Niendam, T. A., Laird, A. R., Ray, K. L., Dean, Y. M., Glahn, D. C., & Carter, C. S. (2012).

Meta-analytic evidence for a superordinate cognitive control network subserving diverse executive functions. *Cognitive, Affective, & Behavioral Neuroscience*, 12(2), 241–268.

[75] Odegaard, B., Knight, R. T., & Lau, H. (2017). Should a few null findings falsify prefrontal theories of conscious perception? *Journal of Neuroscience*, 37(40), 9593–9602.

[76] Olshausen, B. A., & Field, D. J. (1996). Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature*, 381, 607–609.

[77] O'Reilly, R. C., & Frank, M. J. (2006). Making working memory work: A computational model of learning in the prefrontal cortex and basal ganglia. *Neural Computation*, 18(2), 283–328.

[78] Passingham, R. E. (1993). *The frontal lobes and voluntary action*. Oxford, UK: Oxford University Press.

[79] Passingham, R. E., & Wise, S. P. (2012). *The neurobiology of the prefrontal cortex: Anatomy, evolution, and the origin of insight*. Oxford, UK: Oxford University Press.

[80] Petrides, M. (2000). The role of the mid-dorsolateral prefrontal cortex in working memory. *Experimental Brain Research*, 133(1), 44–54.

[81] Petrides, M., & Pandya, D. N. (2001). Dorsolateral prefrontal cortex: Comparative cytoarchitectonic analysis in the human and the macaque brain and corticocortical connection patterns. *European Journal of Neuroscience*, 11(3), 1011–1036.

[82] Poldrack, R. A. (2011). Inferring mental states from neuroimaging data: From reverse inference to large-scale decoding. *Neuron*, 72(5), 692–697.

[83] Postle, B. R. (2006). Working memory as an emergent property of the mind and brain. *Neuroscience*, 139(1), 23–38.

[84] Purves, D., Riddle, D. R., & LaMantia, A.-S. (1992). Iterated patterns of brain circuitry (or how the cortex gets its spots). *Trends in Neurosciences*, 15(10), 362–368.

[85] Rainer, G., Asaad, W. F., & Miller, E. K. (1998). Selective representation of relevant information by neurons in the primate prefrontal cortex. *Nature*, 393, 577–579.

[86] Rao, S. C., Gregor, R., & Miller, E. K. (1997). Integration of what and where in the primate prefrontal cortex. *Science*, 276, 821–824.

[87] Reynolds, J. H., & Chelazzi, L. (2004). Attentional modulation of visual processing. *Annual Review of Neuroscience*, 27(1), 611–647.

[88] Riggall, A. C., & Postle, B. R. (2012). The relationship between working memory storage and elevated activity as measured with functional magnetic resonance imaging. *Journal of Neuroscience*, 32(38), 12990–12998.

[89] Rigotti, M., Barak, O., Warden, M. R., Wang, X. J., Daw, N. D., Miller, E. K., & Fusi, S. (2013). The importance of mixed selectivity in complex cognitive tasks. *Nature*, 497, 585–590.

[90] Rigotti, M., Ben Dayan Rubin, D. D., Wang, X.-J., & Fusi, S. (2010). Internal representation of task rules by recurrent dynamics: The importance of the diversity of neural responses. *Frontiers in Computational Neuroscience*, 4. doi: 10.3389/fncom.2010.00024

[91] Rossi, A. F., Pessoa, L., Desimone, R., & Ungerleider, L. G. (2009). The prefrontal cortex and the executive control of attention. *Experimental Brain Research*, 192(3), 489–497.

[92] Rougier, N. P., Noelle, D. C., Braver, T. S., Cohen, J. D., & O'Reilly, R. C. (2005). Prefrontal cortex and flexible cognitive control: Rules without symbols. *Proceedings of the National Academy of Sciences*, 102(20), 7338–7343.

[93] Sakai, K. (2008). Task set and prefrontal cortex. *Annual Review of Neuroscience*, 31(1), 219–245.

[94] Shenhav, A., Botvinick, M. M., & Cohen, J. D. (2013). The expected value of control: An integrative theory of anterior cingulate cortex function. *Neuron*, 79(2), 217–240.

[95] Sherrington, C. S. (1906). Observations on the scratch-reflex in the spinal dog. *Journal of Physiology*, 34(1–2), 1–50.

[96] Siegel, M., Buschman, T. J., & Miller, E. K. (2015). Cortical information flow during flexible sensorimotor decisions. *Science*, 348, 1352–1355.

[97] Spearman, C. (1904). "General intelligence," objectively determined and measured. *The American Journal of Psychology*, 15(2), 201–293. https://doi.org/10.2307/1412107

[98] Stout, D., Chaminade, T., Thomik, A., Apel, J., & Faisal, A. (2018). Grammars of action in human behavior and evolution. *BioRxiv*. doi: 10.1101/281543

[99] Stroop, J. R. (1935). Studies of interference in serial verbal reactions. *Journal of Experimental Psychology*, 18(6), 643–662. https://doi.org/10/b77m95

[100] Stuss, D. T., & Benson, D. F. (1984). Neuropsychological studies of the frontal lobes. *Psychological Bulletin*, 95(1), 3–28.

[101] Tsujimoto, S., & Postle, B. R. (2012). The prefrontal cortex and oculomotor delayed response: A reconsideration of the "mnemonic scotoma." *Journal of Cognitive Neuroscience*, 24(3), 627–635.

[102] Veniero, D., Gross, J., Morand, S., Duecker, F., Sack, A. T., & Thut, G. (2021). Top-down control of visual cortex by the frontal eye fields through oscillatory realignment. *Nature Communications*, 12(1). doi: 10.1038/s41467-021-21979-7

[103] Voloh, B., Eisenreich, B. R., Maisson, D. J.-N., Ebitz, R. B., Park, H. S., Hayden, B. Y., & Zimmermann, J. (2021). Hierarchical organization of rhesus macaque behavior. *bioRxiv*. doi:

10.1101/2021.11.15.468721

[104] Wallis, J. D., Anderson, K. C., & Miller, E. K. (2001). Single neurons in prefrontal cortex encode abstract rules. *Nature*, 411, 953–956.

[105] Wallis, J. D., & Miller, E. K. (2003). From rule to response: neuronal processes in the premotor and prefrontal cortex. *Journal of Neurophysiology*, 90(3), 1790–1806.

[106] Wang, X.-J. (2002). Probabilistic decision making by slow reverberation in cortical circuits. *Neuron*, 36(5), 955–968.

[107] Warden, M. R., & Miller, E. K. (2010). Task-dependent changes in short-term memory in the prefrontal cortex. *Journal of Neuroscience*, 30(47), 15801–15810.

[108] White, I. M., & Wise, S. P. (1999). Rule-dependent neuronal activity in the prefrontal cortex. *Experimental Brain Research*, 126(3), 315–335.

[109] Wise, S. P., Murray, E. A., & Gerfen, C. R. (1996). The frontal cortex-basal ganglia system in primates. *Critical Reviews in Neurobiology*, 10(3-4), 317–356.

[110] Wood, J. N., & Grafman, J. (2003). Human prefrontal cortex: Processing and representational perspectives. *Nature Reviews Neuroscience*, 4(2), 139–147.

[111] Woolgar, A., Jackson, J., & Duncan, J. (2016). Coding of visual, auditory, rule, and response information in the brain: 10 Years of multivoxel pattern analysis. *Journal of Cognitive Neuroscience*, 28(10), 1433–1454.

[112] Xu, Y. (2017). Reevaluating the sensory account of visual working memory storage. *Trends in Cognitive Sciences*, 21(10), 794–815.

[113] Yang, G. R., Cole, M. W., & Rajan, K. (2019a). How to study the neural mechanisms of multiple tasks. *Current Opinion in Behavioral Sciences*, 29, 134–143.

[114] Yang, G. R., Joglekar, M. R., Song, H. F., Newsome, W. T., & Wang, X.-J. (2019b). Task representations in neural networks trained to perform many cognitive tasks. *Nature Neuroscience*, 22(2), 297–306.

[115] Yeung, N., Nystrom, L. E., Aronson, J. A., & Cohen, J. D. (2006). Between-task competition and cognitive control in task switching. *Journal of Neuroscience*, 26(5), 1429–1438.

[116] Zhou, H., & Desimone, R. (2011). Feature-based attention in the frontal eye field and area V4 during visual search. *Neuron*, 70(6), 1205–1217.