4 Computational Models of Cognitive Control: Past and Current Approaches

Debbie M. Yee and Todd S. Braver
Washington University in St. Louis

4.1. Introduction

A core challenge of cognitive, computational, and systems neuroscience research is to provide a satisfying answer to the following question: how does cognition arise from neural systems? Although researchers have spent decades using a variety of tools (e.g., magnetic resonance imaging, electroencephalography, single-unit recordings) to investigate this question, we have only begun to scratch its surface in terms of understanding how neural substrates work together in synchrony to give rise to complex cognitive processes.

To provide an analogy, imagine listening to a concerto performed by a symphony orchestra. Perhaps you are interested in understanding how the orchestra can blend together so many different sounds from vastly different instruments to give rise to this beautiful masterpiece. In the initial hearing, the piece sounds clearly melodic, lyrical, and filled with multiple complex musical layers that sound cohesive when in concert. However, upon closer examination, it becomes evident that even such complex musical layers can be deconstructed into the contributions from different instruments within the entire ensemble. One approach for understanding the concerto may be simply to listen to one instrument or one section (e.g., attending to a violin solo or the entire violin section when playing the same melody); however, that would only provide a small window into how that specific instrument contributes to the entire piece. Another approach would be to parse out all of the sounds in the piece by instrument, which provides a structural division of the different sounds that comprise the concerto, but neglects the temporal ordering of when the instruments are played, an important aspect of the composition. Perhaps the most insidious problem is that even if we are able to understand the structural and temporal aspects of how each instrument contributes to this specific concerto, the same instruments in this symphony orchestra can also perform a wide variety of other compositions (e.g., other concertos, sonatas, ballads) at other periods in time! Thus, the characterization of the violin's contribution

to the current concerto may not be applicable when considering other musical performances, which makes this type of analysis effort not quite as generalizable as one might have hoped.

4.1.1 The Homunculus Problem of Cognitive Control

The challenge of this problem and the "orchestra concerto" metaphor becomes particularly salient when considering one of the most compelling mysteries of human cognition: how the brain enables human beings to plan, implement, and accomplish the types of controlled, complex, and temporally extended goal-directed behaviors that make up much of modern daily life (e.g., preparing a multi-course meal, constructing IKEA furniture from an instruction manual, writing a computer program, solving a Sudoku puzzle, or figuring out how to successfully complete an MD or PhD). In the orchestra metaphor, it would be akin to understanding how the conductor guides the ensemble to put together a beautiful-sounding and cohesive concerto performance. This mystery has often been posed as the "homunculus problem," which presents the following conundrum: if control over thoughts and action emerges from brain function, then are there special neurobiological and computational properties that differentiate the components that should be labeled as "controller" from the components that are "controlled"? Does the controller/controlled distinction even make sense? And if not, how are we ever going to understand the emergence of intelligent, goal-directed behavior in neurobiological terms?

Within psychology and neuroscience, researchers have often taken a primarily localizationist approach, studying individual brain regions in terms of their associated cognitive functions (Poldrack 2007). At the other extreme is the integrationist perspective, which focuses on the entire brain, parsing it into networks that may be structurally or functionally related (Eliasmith et al. 2012). However, neither of these approaches has yet provided a fully satisfying answer to the fundamental problem of cognitive control. Indeed, as this discussion hopefully makes clear, properly addressing the seemingly intractable homunculus problem likely requires a computational modeling approach. Computational approaches can be utilized in both a reductionist and emergentist manner: deconstructing the mysterious intelligence of the homunculus into hopefully more understandable "dumb" neural subcomponents, while at the same time making clear how complex control functions can emerge from the dynamic interactions among these multiple, simpler subcomponents of cognitive control.

Computational modeling approaches to cognitive control are uniquely powerful, relative to other neuroscience techniques, in that they provide the researcher with a means of generating specific and concrete hypotheses, along with explicit experimental

predictions regarding generative and causally efficacious control mechanisms and their influence on brain activity and behavior (Botvinick and Cohen 2014; O'Reilly 2006; O'Reilly, Herd, and Pauli 2010). More broadly, within the cognitive sciences, the utility of modeling approaches has long been established and appreciated (Newell and Simon 1961). Over thirty years ago, and as described in chapter 1 and figure 1.4, David Marr attempted to formalize these approaches by articulating an influential proposal for decomposing and investigating complex cognitive systems across three levels of analysis: the *computational*, the *algorithmic*, and the *implementational* (Marr 1982; Bechtel 1994). These levels of analyses were initially introduced to tackle computational questions in vision, and have been criticized by various researchers as potentially being too rigid to be universally applicable (Dayan 2006). Yet the Marr framework can be fruitfully applied when considering complementary questions about the neural and computational mechanisms that underlie more complex temporally extended goal-directed behavior, such as: What computational goal is accomplished by a putative control function? What is the algorithm that encodes this function? Can we identify the neural systems and mechanisms that implement the algorithm? Consequently, we will make use of the Marr framework in this chapter, in order to provide a general intuition for how various computational models attempt to address specific questions about cognitive control function.

4.1.2 Why Cognitive Control?

The current chapter highlights past and current computational models of cognitive control, and the purpose is twofold. First, cognitive control is a well-known psychological construct, with a long history of researchers using computational modeling approaches to attempt to explain its underlying cognitive mechanisms (Newell and Simon 1972; Rumelhart et al. 1986; Cohen, Dunbar, and McClelland 1990; Braver and Cohen 2000; Anderson et al. 2008). Second, cognitive control ability is disrupted across a wide range of mental disorders, with a vast body of literature now supporting the hypothesis that cognitive control impairments are prominent in many such disorders, including schizophrenia, depression, obsessive-compulsive disorder, ADHD, addiction, Alzheimer's disease, and Parkinson's disease (Lesh et al. 2011; Fales et al. 2008; Halari et al. 2009; Greisberg and McKay 2003; van Meel et al. 2007; Vaidya et al. 2005; Belleville, Chertkow, and Gauthier 2007; Brown and Marsden 1990; Wylie et al. 2010; Snyder, Miyake, and Hankin 2015). Indeed, it may not be an exaggeration to argue that an impairment of cognitive control, in one form or another, is the defining feature of many forms of mental illness. Thus, understanding the mechanisms that underlie cognitive control function can provide a crucial window into psychopathology.

Cognitive control is operationalized as the ability to perform task-relevant processing in the face of distractions or in the absence of environmental support, specifically by active maintenance and flexible updating of task representations over time, in order to pursue task-relevant objectives and behavioral goals (Engle and Kane 2004; Braver 2012; O'Reilly, Braver, and Cohen 1999). A core tenet of cognitive control is the distinction between controlled and automatic processing (Posner and Snyder 1975; Shiffrin and Schneider 1977; Norman and Shallice 1986). It is now generally appreciated that a fundamental tradeoff exists between recruiting and directing cognitive resources to deliberately perform a demanding task versus carrying out less effortful and habitual responses that may require fewer attentional resources, but that also may be less flexible. Typically, the allocation of control depends on the amount of cognitive effort or mental demand required. In other words, the control of behavior arises from the cognitive demands imposed by the requirement to successfully perform a task, and effort allocation arises from the dynamic recruitment of available cognitive processes that can appropriately meet these demands during task performance (Botvinick and Cohen 2014). Some have proposed various computational models and frameworks to understand this tradeoff between effort and automaticity in controlled behavior (Cohen, Dunbar, and McClelland 1990; Schneider and Chein 2003), whereas others have hypothesized that humans perform cost-benefit analyses between expected payoff and cognitive effort to determine the optimal allocation of cognitive control (Shenhav et al. 2017; Dixon and Christoff 2012; Kool and Botvinick 2014; Westbrook and Braver 2015). All in all, there still remain many unanswered questions regarding the computational and neural mechanisms that underlie cognitive control; we argue these can be more adequately addressed with computational modeling approaches.

As a brief aside, we wish to acknowledge that such computational modeling approaches have been prevalent and successful in advancing understanding for other related, but potentially more specialized higher cognitive processes, such as attention (Gershman, Cohen, and Niv 2010), learning (Tenenbaum, Griffiths, and Kemp 2006), semantic knowledge (Rogers and McClelland 2004), and memory (Polyn, Norman, and Kahana 2009). Thus, while this chapter will focus primarily on cognitive control, we hope that the reader may extrapolate these principles to obtain a broader perspective for how computational models can be used to study other cognitive systems.

4.1.3 Roadmap to This Chapter

This chapter contains two main sections. First, we will provide a brief review of several key computational models that have been influential in advancing understanding of cognitive control mechanisms. This review of such models is not meant to

be comprehensive but will hopefully provide a useful primer for readers to become familiar with classical and current computational models of cognitive control, with the understanding that the principles behind these models can be extended to other related models. Next, we discuss key features of computational models that make them particularly useful and generative in guiding further research efforts (i.e., what "tests" can we run to determine whether a computational model can make accurate and generalized predictions about controlled behavior?). The chapter concludes with a concrete example of how such modeling frameworks can be used to make predictions in mental illness, with some speculation about how cognitive control function breaks down in schizophrenia, a psychiatric disorder hypothesized to be strongly associated with cognitive control impairment.

4.2 Past and Current Models of Cognitive Control

A broad range of computational models have played a prominent role in the development and understanding of cognitive control theory and its underlying mechanisms, including those that have primarily arisen from symbolic modeling traditions, such as those involving production system architectures (ACT-R, Anderson 1996; EPIC, Kieras and Meyer 1997). At the other end of the spectrum are models arising from the computational neuroscience tradition (Wang 2013), similar to those covered in chapter 3. Here, we focus on four contemporary models that address challenging and unique computational problems integral to cognitive control function, and which have also played an influential role in advancing research within this domain:

1. How do we determine when to actively maintain versus rapidly update contextual information in working memory?
2. How is the demand for cognitive control evaluated, and what is the computational role of the anterior cingulate cortex?
3. How do contextual representations guide action selection during hierarchically organized task goals, and what is the computational role of the prefrontal cortex (PFC)?
4. How are task sets learned and organized during behavioral performance, and when do they generalize to novel contexts?

4.2.1 How Do We Determine When to Actively Maintain versus Rapidly Update Contextual Information in Working Memory?

A key cognitive control challenge is in determining what information is relevant to be maintained (i.e., in working memory) during the pursuit of task goals, and when this

information should be updated with newer task-relevant information. A potentially useful analogy for visualizing this issue is the concept of a "mental blackboard," which describes the dilemma of deciding between when learned information in working memory should be kept, or instead erased and overwritten (Baddeley 1986). Early computational models attempted to use attractor models to understand the mechanisms that underlie robust active maintenance of working memory against irrelevant distractors (Changeux and Dehaene 1989; Zipser et al. 1993; Cohen, Braver, and O'Reilly 1996; Compte et al. 2000; Durstewitz, Seamans, and Sejnowski 2000; Deco and Rolls 2003). However, a major limitation of these models is their lack of a mechanism for precisely updating working memory when newer, task-relevant information is introduced. This tension between these two working memory functions is difficult to reconcile, as they inherently contradict each other—active maintenance increases resistance to distractors, whereas flexible updating makes the system more vulnerable to distraction. Thus, the computational challenge lies in building a model that can explain how a system regulates the fundamental trade-off between learning when to actively maintain context representations (i.e., task-relevant information that is internally represented) to achieve controlled processing versus rapidly updating new information into working memory, a core problem of cognitive control (O'Reilly, Braver, and Cohen 1999; Braver and Cohen 2000).

One approach toward understanding the computational mechanisms that underlie this trade-off comes from the "parallel-distributed-processing" approach (also dubbed "connectionist" or "neural network" models in the literature; see section 2.1). These models view control as arising from the interaction of multiple relatively simple elements (e.g., neurons or neural assemblies that perform local processes within a single brain system or unit). Thus, the models emphasize how cognitive control functions emerge from a network of brain regions activated interactively and in parallel, rather than the more historical modular approach of localizing cognitive function to a single brain region (Hinton 1984; O'Reilly 2006).

A well-established model from within the connectionist tradition is the PFC and basal ganglia (PBWM) model developed by Frank, O'Reilly, and their colleagues (Frank, Loughry, and O'Reilly 2001; O'Reilly and Frank 2006; Hazy, Frank, and O'Reilly 2007). In the PBWM model, the PFC and basal ganglia (BG) interact to solve the maintenance versus updating problem by implementing a flexible working memory system with an adaptive gating mechanism. This represents an elegant algorithmic solution for resolving this computational question, as it provides two separate modes of working memory that optimize active maintenance and flexible updating, respectively (figure 4.1a). Specifically, working memory is insulated from distractor signals (i.e., irrelevant sensory

input) when the gating mechanism is closed, but is receptive to utilizing information from such sensory signals when gating mechanisms are open. However, the introduction of this gating mechanism then begs the following question: how does the brain know when to open or close the gate? In other words, who or what controls the gate?

At the biological (i.e., implementational) level, the PBWM model proposes that the PFC facilitates the active maintenance mechanisms for sustaining task-relevant information, whereas the BG provides the selective gating mechanism, which independently switches between updating versus maintenance of information in PFC. Specifically, the key component of PBWM is that the BG performs this selective dynamic gating via disinhibition and, moreover, that this dynamic gating functionality depends upon the dopaminergic system (DA, see figure 4.1b). In this framework, dopaminergic "Go" neurons in dorsal striatum fire to disinhibit PFC to enable updating of working memory representations in PFC, while "NoGo" neurons counteract this effect to support robust maintenance of PFC working memory representations and resistance to distractions.

Notably, other computational models have proposed similar gating mechanisms that regulate flexible updating and maintenance of task-relevant representations during working memory, but driven primarily by direct dopamine (DA) projections to PFC (Braver and Cohen 1999, 2000). However, a criticism of the global DA-firing hypothesis is that this mechanism would not fully explain more complex cognitive tasks in which individuals would need to maintain and update different task representations simultaneously, such as when there is a hierarchical structure to working memory (e.g., remembering to press a button for a specific stimulus only during context A, but not context B).

Taken together, the PBWM leverages the gating mechanism as an algorithmic solution to the computational problem of switching between active maintenance and flexible updating within working memory mechanisms. This model suggests that the PFC implements active maintenance of task-relevant information, whereas the BG contains selective gating mechanisms which switch between "robust maintenance" and "selective updating" of information held in PFC during working memory. Midbrain DA release is hypothesized to modulate this gating mechanism. However, exactly how, when, and where DA firing drives these working memory functions (e.g., only in the BG or also directly in PFC), is a question that remains to be fully explored.

4.2.2 How Is the Demand for Cognitive Control Evaluated, and What Is the Computational Role of the Anterior Cingulate Cortex?

Another core computational challenge within the domain of cognitive control is the following: how is the current demand for control evaluated, and in what form is this

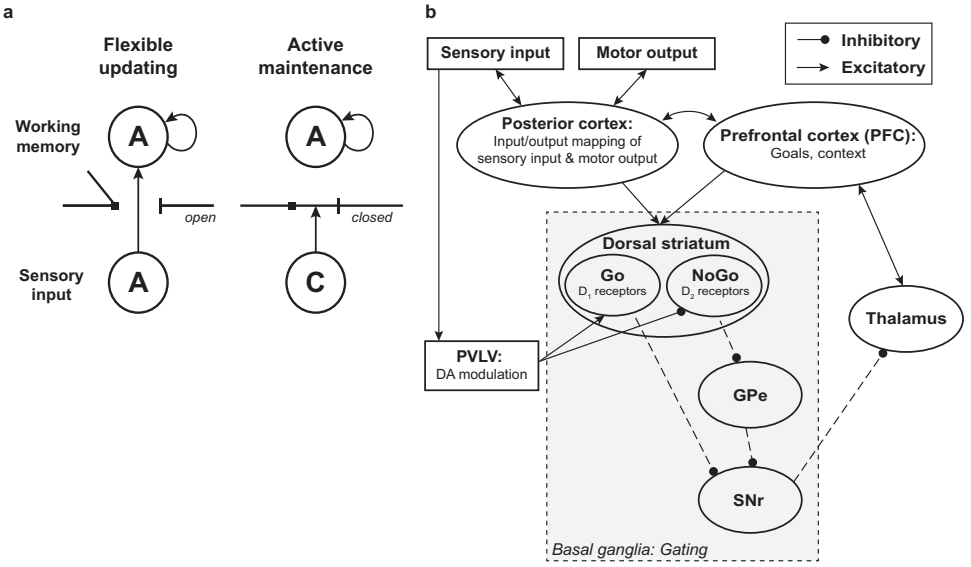

Figure 4.1

A) Gating mechanism from Frank and O'Reilly's prefrontal basal ganglia and working memory (PBWM) model (Frank, Loughry, and O'Reilly 2001). At the algorithmic level, this connectionist computational model features a gating function, which switches between active maintenance and flexible updating of working memory to incorporate task-relevant information, two core functions of cognitive control. B) Neural network model implementation of the PBWM. Here, sensory inputs are mapped onto motor outputs via posterior ("hidden") layers. The PFC contextualizes this information and encodes relevant prior information and goals. The basal ganglia (BG) updates the PFC via dynamic gating, which is driven by dopaminergic modulation from a separate "PVLV" (primary value and learned value learning algorithm) system (O'Reilly et al. 2007). Specifically, dopamine (DA) is excitatory onto the Go neurons via D_1 receptors and inhibitory onto NoGo neurons via D_2 receptors. Thus, increased DA firing will inhibit SNr (substantia nigra pars reticulata) and disinhibit PFC to facilitate flexible updating of working memory representations in PFC. Decreased DA firing, on the other hand, counteracts this effect and facilitates active maintenance of current working memory representations in PFC.

evaluative signal transmitted? In other words, how does the brain determine which situations or task conditions require more mental resources (than are currently available) to successfully pursue task goals, and what is the necessary relevant information that underlies this evaluation? This type of question is difficult to address from a purely theoretical perspective, as "cognitive demand" is an elusive construct that appears to arise under a wide variety of mentally challenging tasks. Thus, a prerequisite for building a computational solution is understanding which experimental conditions demand and elicit greater cognitive control, along with identifying relevant behavioral measures as empirical evidence for increased cognitive effort (note that in the literature, the terms cognitive effort and mental effort are used interchangeably).

A plethora of work has identified tasks with behavioral measures that demonstrate selective recruitment of cognitive control (Botvinick, Cohen, and Carter 2004; Ridderinkhof et al. 2004; Braver and Ruge 2006). For example, in the Stroop task, cognitive control is required to override the prepotent response to read a word, in order to perform the correct task of reading the color ink of the word. In the N-back, cognitive control is required to respond selectively to N-back matches (e.g., in a two-back task, a target response should be given only if the current stimulus matches the one presented two stimuli ago) rather than based on simple familiarity. In the stop-signal (or change-signal) task, cognitive control is required to cancel an already initiated behavioral response if a stop signal (or change cue) is presented. In the Erikson flanker task, cognitive control is required to respond selectively to a centrally presented stimulus and ignore the flanker stimuli, particularly when these are distracting and incongruent with the central stimulus. Critically, all of these tasks contain experimental conditions that reliably increase cognitive control demands in a transient, trial-by-trial manner (i.e., the cognitive system monitors ongoing responses and adjusts to the level of cognitive control needed on the current trial). Likewise, they are indexed by specific behavioral measures that reflect this enhanced cognitive control demand (e.g., Stroop interference effect, stop-signal reaction time).

A well-established finding is that canonical control tasks, such as the ones listed above, consistently co-activate the dorsolateral PFC (dlPFC) and the dorsomedial PFC (Egner 2009; Duverne and Koechlin 2017), a brain region that spans the dorsal anterior cingulate cortex (ACC) and presupplementary motor area (Duncan and Owen 2000; Duncan 2010). The dlPFC is thought to play a primary role in actively maintaining representations of task goals and the associated actions (or behavioral rules) needed to achieve them. In contrast, the ACC is thought to be involved in signaling when more control should be implemented by the dlPFC to accomplish these goals. It is generally accepted that the interaction between these two brain regions

is important for dynamically adjusting cognitive control. Many have argued for the ACC as an important locus of cognitive control (Holroyd et al. 2004; Kerns 2004), although there remains much controversy over what actual information is represented by the ACC and signaled to the dlPFC to indicate that cognitive control is needed during tasks.

Several prominent theoretical accounts of ACC's computational role in cognitive control have arisen in recent years, including the detection of error signals (Gehring et al. 1993; Holroyd et al. 2005), reinforcement learning (Holroyd and Coles 2002), conflict monitoring (Botvinick et al. 2001; Botvinick, Cohen, and Carter 2004), error likelihood (Carter et al. 1998; Brown and Braver 2005), cost-benefit analyses of implementing control (Shenhav, Botvinick, and Cohen 2013), and even uncertainty in the environment (Behrens et al. 2007). An account developed to reconcile and unify these divergent perspectives is the predicted response-outcome (PRO) model (figure 4.2; Alexander and Brown 2011, 2014). The PRO model contains two components. One component of the model learns to predict multiple likely outcomes of various chosen actions, regardless of whether these outcomes are good or bad (i.e., response-outcome learning). A second component of the model detects discrepancies between actual and predicted outcomes and uses this prediction-error signal (i.e., actual outcomes minus expected outcomes) to update and refine subsequent predictions. Moreover, a key aspect of the prediction-error signal is that it also indicates "negative surprise," when an expected outcome does not occur. This form of negative surprise signal can indicate not only when an unexpected error occurs, but also when the response is slower than expected or when the correct action is more ambiguous (which is likely to happen on trials associated with high response conflict).

At the implementational level, the PRO model postulates that separate neural signals within ACC represent outcome prediction and prediction error (negative surprise), respectively. Specifically, the model suggests that the prediction signal should reliably increase immediately prior to when the most likely outcome will occur (i.e., a pre-response anticipatory signal). The negative surprise signal, on the other hand, will reliably activate after the action that produces an unpredicted outcome has occurred (i.e., a post-response evaluative signal). Critically, these hypotheses have been tested empirically across multiple tasks (e.g., change-signal task, Erikson flanker), as well as across different types of neural data (e.g., fMRI activity, event-related potentials, monkey single-unit neurophysiology). This validation of the PRO model across such a wide range of neural data demonstrates that it provides a useful generalizable computational algorithm by which the ACC can signal an increased need for cognitive control. Recent

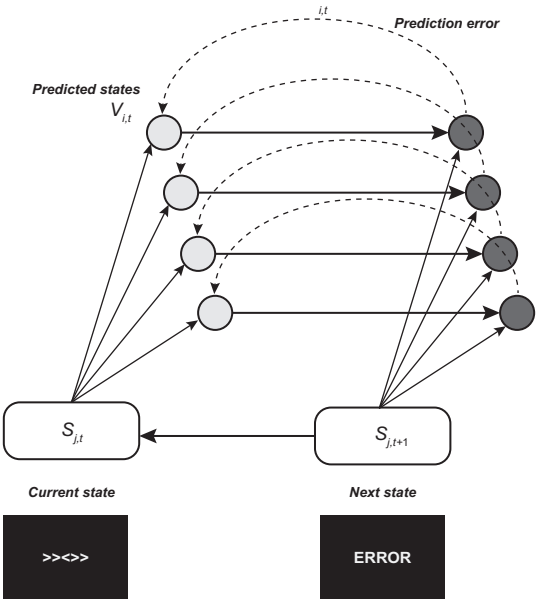

Figure 4.2
Schematic of the predicted response-outcome (PRO) model by Alexander and Brown (2011, 2014). First, the PRO model learns predictions of multiple possible future outcomes of various chosen actions (indicated by $V_{i,t}$), using an error-likelihood signal. Thus, activity in the PRO model reflects a temporally discounted prediction of such outcomes, which are proportionate to their likelihood of occurrence. Second, the PRO detects discrepancies between predicted and observed outcomes, and then uses their prediction-error signal (δ) to update and improve subsequent predictions. *S* refers to the representation of the stimulus (e.g., conflicting arrows from the Erikson flanker task) or task-related feedback (e.g., a screen indicating an error was made). Thus, the PRO model continually learns and updates associations between task-related cues and feedback in cognitive tasks.

efforts have attempted to expand this account to include hierarchical representation within ACC and dlPFC (Alexander and Brown 2015), a topic relevant to the next section. Other recent efforts have attempted to link ACC signals with more affective/motivational quantities (Vassena, Holroyd, and Alexander 2017). These include the expected value of control (EVC; Shenhav, Botvinick, and Cohen 2013) and related accounts (Holroyd and McClure 2015; Westbrook and Braver 2016), which postulate that ACC regulates the allocation and persistence of cognitive effort based on signals indicating the current subjective motivational (and/or hedonic) value of task and goal outcomes.

4.2.3. How Do Contextual Representations Guide Action Selection toward Hierarchically Organized Task Goals, and What Is the Computational Role of the Prefrontal Cortex?

A third computational question of control relates to the issue of abstraction. How can a "high-level" goal constrain and implement a "lower-level" goal? As an example, imagine the following scenario: you hear a nearby phone ring, and you have an instinctive impulse to answer it. However, context plays an important role in your action plan, so while you might automatically answer a nearby phone in your own home, you would inhibit this tendency to answer a ringing phone at your friend's home. Yet you might switch your action plan if your preoccupied friend asks you to answer the ringing phone on their behalf (e.g., when they are busy with a task). This example articulates a fundamental computational challenge of implementing task goals—specifically, how do humans utilize contextual representations and higher-level goals to guide action selection during pursuit of lower-level goals, and how does the brain implement this type of hierarchical control?

One promising algorithmic solution for this perplexing question is the concept of hierarchical organization of task–goal representations. The notion of applying hierarchical structure to parse complex systems into subordinate and interrelated subsystems has long been established, with subsystems being further subdivided into "elementary" units (Simon 1962). Similarly, some theorists have argued that control signals used to guide behavioral actions, based on internal plans and goals, can also be subdivided into sensorimotor, contextual, and episodic levels of control (Koechlin, Ody, and Kouneiher 2003; Koechlin and Summerfield 2007; figure 4.3). Critically, this information-theoretic model (i.e., based on principles from information theory; Shannon 1948), which has also been termed the "cascade model," postulates that the hierarchical division occurs according to a temporal dimension; that is, when in time control is implemented. Specifically, according to the model, actions selected based on temporally proximal stimuli would be lower on the hierarchy, whereas actions selected based on past information that is actively maintained in conjunction with the recent stimulus would be higher on the hierarchy. According to this framework, greater demand for cognitive control can also be formalized as the amount of information required to be actively maintained over longer time periods to enable successful behavioral action selection. As a brief aside, it is worth noting that earlier models also utilized hierarchical frameworks to understand temporal abstraction in behavior (Cooper and Shallice 2006), but the primary thrust of the cascade model and related variants has been to use reinforcement learning to subdivide temporally abstract complex action plans (i.e., "options") into simpler behaviors, an adaptive and efficient

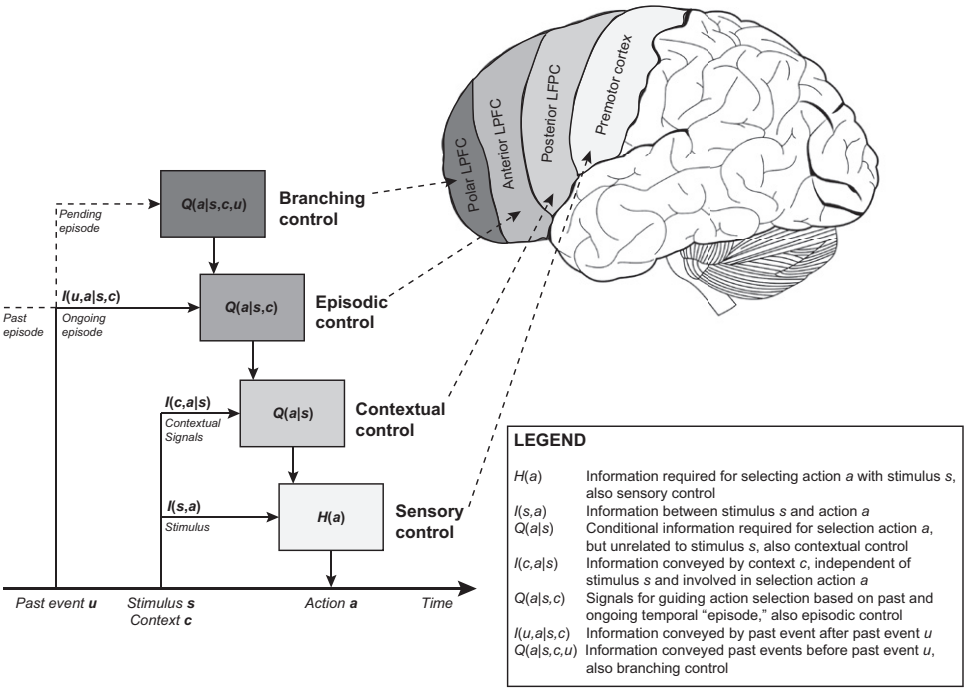

Figure 4.3

Model of hierarchical cognitive control by Koechlin and colleagues (2003, 2007). This information-theoretic model posits that cognitive control operates according to three nested levels of control processes (branching, episodic, contextual), which are implemented as a cascade from anterior to posterior prefrontal regions. $H(a)$ represents sensory control, the information required to select an action (a) to appropriate incoming stimuli, and is the sum of two control terms: bottom-up information conveyed by the stimulus (s) regarding the appropriate action [$I(s, a)$] and top-down information processed in the posterior lateral PFC [$Q(a|s)$]. The Q(a|s) term represents contextual control, the incoming signals congruent with the subject's response, and is the sum of two control terms: bottom-up information from the contextual (c) signals and stimulus [$I(c, a|s)$, $I(s, a)$], and top-down information processed in anterior lateral PFC [$Q(a|s,c)$]. The $Q(a|s,c)$ term represents episodic control, neural signals that guide actions based on information retrieved from past events stored in episodic memory (i.e., tonically maintained over a longer temporal interval), which is the sum of bottom-up information from past event u [I(u,a|s,c)] and top-down information processed in the polar lateral PFC [$Q(a|s,c,u)$]. The branching control term [$Q(a|s,c,u)$] relates to information conveyed by events prior to event u, and are maintained until the current episode or trial is complete. Thus, this computational model parses different levels of control based on how much information must be internally represented and actively maintained in order to select and perform a correct action.

encoding strategy relevant for understanding structured abstract action representations (Botvinick 2008; Botvinick, Niv, and Barto 2009; Solway et al. 2014; Holroyd and Yeung 2011).

At the neural level, the cascade model implements hierarchical cognitive control along the anterior-posterior (i.e., rostral–caudal) axis of lateral PFC, with control signals higher up in the hierarchy represented in more anterior prefrontal regions (Koechlin, Ody, and Kouneiher 2003; Badre 2008; Badre and D'Esposito 2009). Although it is well accepted that PFC subserves high-level cognitive function and cognitive control, researchers have only recently attempted to build a parcellation scheme of this large brain region according to a functional organizing principle (Fuster 2001). Evidence from human neuroimaging studies supports the hypothesis of hierarchical representation, with more anterior regions of lateral PFC being activated when cognitive control is implemented for past information, and posterior regions being activated during action selection from more immediate information (Velanova et al. 2003; Braver and Bongiolatti 2002; Braver, Reynolds, and Donaldson 2003; Badre and D'Esposito 2007; Nee and Brown 2013). Additionally, single-unit studies in nonhuman primates are supportive of the idea that PFC is functionally organized according to the rostral–caudal axis: whereas caudal regions are involved in direct sensorimotor mappings, more rostral regions are involved in higher-order control processes that regulate action selection among multiple competing responses and stimuli (Petrides 2005; Shima et al. 2007). Thus, the hierarchical organization of PFC appears to be central to performing the neural computations underlying task–goal abstraction and action selection. Active research efforts focus on understanding how these divisions in the hierarchy are initially learned (Reynolds and O'Reilly 2009; Frank and Badre 2012), and whether the hierarchical structure is primarily anatomic or dynamic (Reynolds et al. 2012; Nee and D'Esposito 2016).

4.2.4 How Are Task Sets Learned during Behavioral Performance, and When Are They Applied to Novel Contexts?

The fourth and final computational question in this chapter relates to the interaction of cognitive control and learning. In daily life, humans are faced with the challenge of learning a set of actions, sometimes simple or complex, in order to complete a specific task (i.e., a task set). A related challenge is discerning between knowing when task-set rules that are learned in one context can be applied to a novel context (i.e., they generalize), or instead when a new task set needs to be constructed. For example, when searching for the restroom at a shopping mall, one may learn a rule to look for signs

that contain the text "Bathroom" with arrows pointing to a particular location. However, while this task-set rule may be pertinent when navigating malls in the United States, the same strategy may not be effective when searching for a restroom in other countries (e.g., United Kingdom), since the signs may read "W.C." instead of "Bathroom." Broadly speaking, creating a set of behavioral tools not tied to the context in which they were learned is useful, as this strategy enables flexible and efficient learning of task-set rules that can be generalized to novel contexts. However, the neural computations that underlie how cognitive control is deployed to learn task sets are less well understood. Thus, the main motivating computational question is the following: in a new context requiring representation of tasks and task-set rules, is it more effective and efficient to generalize from an existing task-set representation (presumably stably encoded in long-term memory), or to instead build a new representation that is more optimized for the current context?

In the last decade, many accounts of cognitive control looked to algorithms and approaches from the reinforcement learning literature for inspiration in how task-set and goal representations might be acquired (Botvinick, Niv, and Barto 2009; Dayan 2012c). A recent model that directly targeted this learning question is the context-task-set (C-TS) model, which aims to approximate how humans create, build, and cluster task-set structures (Collins and Frank 2013; figure 4.4). The model's algorithm harnesses the power of both reinforcement learning and Bayesian generative processes that can infer the presence of latent states. Specifically, the model is designed to accomplish three goals: 1) create representations of task sets and their parameters; 2) infer at each trial or time point which task set is relevant in order to guide action selection; and 3) discover hidden task-set rules not already in its repertoire. A key element that drives the learning process is context—here defined as a higher-order factor associated with a lower-level stimulus—which influences which action/motor plan would be selected. When the model is exposed to a novel context, the likelihood of selecting an existing task set is based on the popularity of that task set; that is, its relevance across multiple other contexts. Conversely, the probability of creating a new task set is set to be inversely proportional to a parameter indicating conservativeness; that is, the prior probability that the stimulus-action relationship would be governed by an existing rule rather than a new one. Further, if a new task set is created, the model must learn predicted reward outcomes following action selection in response to the current stimulus, as well as determine if the task set is valid for the given context. If a selected action leads to a rewarding outcome, the model then updates the parameters to strengthen the association between a context and a specific task set. Thus, the C-TS model provides

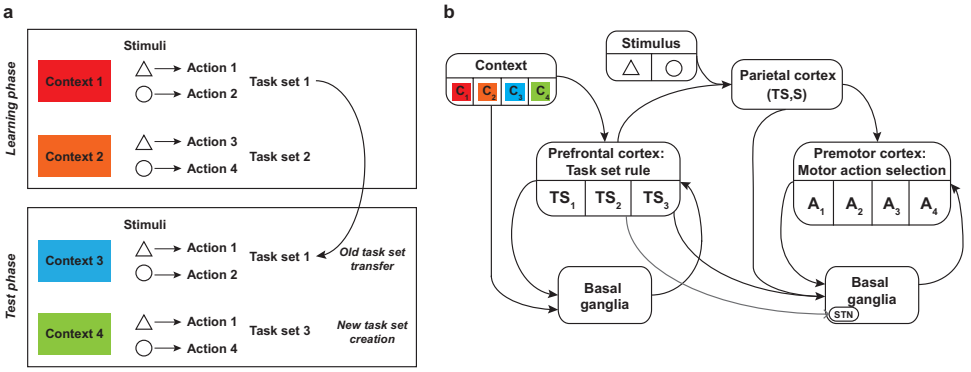

Figure 4.4

(a) Context-task-set (C-TS) model by Collins and Frank (2013). This model solves the problem of how to learn hidden task-set rules (i.e., when in a given state and presented with sensory input, which action should be taken in order to maximize reward). The C-TS model posits that states are determined hierarchically; that is, an agent will consider some input dimension to act as a higher-order context (C), which indicates a task set (TS) and other dimensions to act as lower-level stimuli (S), in determining which motor actions (A) to produce. Here, the color context determines a latent task set that facilitates learning of shape stimulus–action associations in the learning phase (e.g., C_1 is associated with TS_1). In the test phase, C_3 maps onto the same shape stimulus-action association as C_1, so the C_3 context is transferred to TS_1, whereas C_4 should be assigned to a new task set. Critically, the model predicts that it should be faster to transfer a task set than learn a new task set. (b) Schematic of two-loop corticostriatal-gating neural network model. These two loops are nested hierarchically, such that one loop learns to gate an abstract task set (and will group together the contexts that are associated with the same task sets), whereas the other loop learns to gate a motor action response conditioned on the task set and perceptual stimulus. Here, color context (c) serves as the input for learning to select the correct task set in the PFC loop. This information is multiplexed with the shape stimulus in parietal cortex to modulate the motor loop and select the correct motor actions. These two loops accomplish two objectives: 1) constrain motor actions until a task set is selected; and 2) allow conflict at the level of task-set selection to delay responding in the motor loop, preventing premature action selection until a valid task set is selected. Taken together, both algorithmic and neural network models similarly and accurately predict behavioral task performance. The synergism of different modeling levels provides an account of how humans engage cognitive control and learning to produce structured abstract representations that enable generalization in the long term, even if it may be costly in the short term.

a computationally tractable algorithm for task-set learning and clustering that not only feasibly links multiple contexts to the same task set, but also discerns when to build a new task set to accommodate a novel context. This process has been since dubbed "structure learning."

This structure-learning process also has an implementational solution, simulated in a biologically plausible neural network model (in the same PDP tradition as the PBWM model), which provides a specific hypothesis about how structure learning occurs in the brain. In particular, the model formalizes how higher- and lower-level task-set structures and stimulus-action relationships are learned analogously within a distributed brain network involving interactions between PFC and BG. The key functional components of the model are two corticostriatal circuits arranged hierarchically with independent gating mechanisms. The higher-order loop involves anterior regions of PFC and striatum, which learn to gate an abstract task set and cluster contexts associated with the same task set. The lower-order loop between posterior PFC and striatum also projects to the subthalamic nucleus, which provides the capability of gating motor responses based on the selected task set and perceptual stimulus. Thus, the execution of viable motor responses is constrained by task-set selection, and conflict that occurs at the level of task-set selection delays the motor response, thus preventing premature action selection until a valid task set is verified.

Both the algorithmic C-TS and the neural network model lead to similar predictions in human behavior. The convergence between these modeling approaches makes clear their joint utility as explanatory tools for understanding the processes that underlie structure learning. Specifically, together these models make an important claim: that humans have a bias toward structure learning, even when it is costly, because such learning enables longer-term benefits in generalization and overall flexibility in novel situations (Collins 2017).

From a broader perspective, a unique strength of using multiple computational modeling approaches is the ability to provide complementary insight into the cognitive and neural processes that result from the interaction of cognitive control and learning functions. These two variants of the C-TS model provide an admirable exemplar for how to integrate computational, algorithmic, and implemental analysis levels, and thus formalize a theoretical account that can approximate human implementation of cognitive control processing and structure learning. Thus, while the C-TS specifically targets understanding key mechanisms of cognitive control, the multilevel approach adopted to investigate these mechanisms provides excellent scaffolding for future computational investigation in other cognitive research domains.

4.3 Discussion: Evaluating Models of Cognitive Control

Next, we address two relevant issues in evaluating computational models of cognitive control: 1) what are good metrics for determining whether a model provides a useful contribution to our understanding of cognitive control mechanisms? And 2) how can models in this domain be successfully applied to understand the nature of cognitive control deficits in psychiatric disorders?

4.3.1 Model Evaluation: Determining Whether a Computational Model is Useful

A famous adage by the British statistician George E. P. Box states the following—"all models are bad; some models are useful." It is generally accepted that most computational models are limited in their ability to account for all observed behavior, and at best typically encompass the critical data variability within a certain limited cognitive domain (e.g., cognitive control phenomena related to standard experimental response-conflict tasks), but do not generalize well beyond this limited domain, such as to novel tasks or contexts. Another common critique of algorithmic approaches, in particular, is that these computations may not necessarily accurately reflect how cognitive processes are implemented on the biological level. For example, while a model may provide a sufficient hypothesis of cognitive control function and account for the key behavioral variance in a task, it is possible that the brain-behavior relationship may arise from a completely different computational or neural process altogether in the brain. Thus, an important step in this approach is model evaluation; that is, deciding whether a model has utility. In other words, what makes a model useful for advancing cognitive research? Here we describe two complementary metrics for determining the utility of computational models—specifically, examining whether they are descriptive or predictive.

A computational model is *descriptive* if it provides a detailed explanation that accounts for significant variability of observed data (i.e., how well the model fits the data). Since models provide hypotheses about the data-generating process, a descriptive computational model should provide insight into the mechanisms that give rise to the observed behavioral or neural responses in a given task. For example, an indisputable strength of Alexander and Brown's (2011, 2014) PRO model is its ability to account for a diverse range of empirical results, related to evaluation of demands for cognitive control, that span across both human and primate studies. Since the PRO model successfully models diverse neural and behavioral data from multiple cognitive control studies, it consequently provides compelling evidence for the hypothesis that predictive neural computation relating actions to outcomes implemented in the

ACC and associated medial frontal regions may be a useful signal linked to the engagement of cognitive control. However, although the PRO model formalizes one potential algorithmic explanation for the generative process underlying extant data, it may neither reflect the actual neural computations that occur in the brain, nor necessarily accurately predict data outcomes in future studies. Thus, a limitation of this evaluation metric is that while a model with high explanatory power may explain prior data, the proposed mechanism may not be able to explain new data.

Conversely, a computational model is *predictive* if it describes a generative process that accurately forecasts and extrapolates to novel tasks or contexts. A predictive model contains a specific hypothesis about the neural computations that generate relevant data from one task or context and incorporates theory to reliably estimate behavioral and neural outcomes in a novel task/context. Collins and Frank's (2013) convergent C-TS and neural network models provide excellent examples of predictive modeling, as both models make accurate predictions of behavioral outcomes in novel tasks/contexts. Critically, a theoretical assumption guiding development of these models is that humans spontaneously build task-set structure in learning problems. This structure-learning assumption was tested in empirical studies, validating that the model could generalize to task contexts not previously learned. To summarize the key distinction put forth here, both "descriptive" and "predictive" computational models provide process mechanisms for how data are generated, but the former describes how well the model may fit extant data, whereas the latter describes how well the model generalizes to unseen data.

More broadly and generally, a computational model can serve a very useful function if it is explicitly specified to the degree that it can provide a focal point to drive and rejuvenate new research efforts. For example, while there is much controversy over ACC function, computational models have helped to elucidate potentially relevant cognitive mechanisms by providing specific testable hypotheses for empirical study (Botvinick and Cohen 2014; Vassena, Holroyd, and Alexander 2017). Moreover, although models may not always be accurate, they can highlight limitations of existing theory (e.g., what can and cannot be predicted by the model) and provide insight into how the theory should be revised in future iterations. The computational models described in this chapter are theory-driven approaches that attempt to describe how the brain implements cognitive control in an explicit way, in contrast to more vague descriptions by conceptual or verbal models. Thus, by attempting to spell out the exact mechanism for how cognitive control systems can be realized, the models described here provide explicit answers to the mysterious "homunculus" problem of cognitive control. Furthermore, our hope is that such models will eventually be directly useful

for elucidating how and why abnormal psychological and neurological processes arise in mental illness.

4.3.2 Cognitive Control Impairments in Schizophrenia

As an example of the point made above, we conclude this chapter with an example in which computational models of cognitive control have already been directly applied to a psychiatric disorder: specifically, to investigate the etiology of cognitive impairments in schizophrenia. A large literature on cognitive function in schizophrenia has reliably established that patients with this illness demonstrate impairments in attention, working memory, episodic memory, and executive functions (Snitz, MacDonald, and Carter 2006). More specifically, an influential hypothesis is that schizophrenia is characterized by disrupted cognitive control, specifically a disturbance in the ability to internally represent and maintain contextual or task–goal information in the service of exerting control over one's actions or thoughts (Cohen and Servan-Schreiber 1992; Barch and Ceaser 2012; Lesh et al. 2011; Barch, Culbreth, and Sheffield 2018). A key feature of the account is that such disruptions in cognitive control and context representation are directly linked to dysfunction of the DA neuromodulation in PFC, which has long been suggested to be a primary mechanism of pathophysiology in schizophrenia (Meltzer and Stahl 1976; Snyder 1976; Seeman 1987; Toda and Abi-Dargham 2007; Rolls et al. 2008; Valton et al. 2017). In particular, a common view is that at least some of the cognitive impairments observed in schizophrenia putatively are related to reduced dysfunctional DA signaling in striatum and PFC, as well as increased "noise" potentially resulting from increased tonic DA activity or aberrant phasic DA activity (Braver, Barch, and Cohen 1999; Rolls and Grabenhorst 2008; Maia and Frank 2017).

As a direct test for this hypothesis of dysregulated cognitive control and its relationship to DA and PFC, Braver and colleagues modified an extant computational model of PFC function and context processing. Specifically, the goal was to make explicit predictions about behavioral and brain activity patterns that would be observed in schizophrenia patients performing the AX Continuous Performance Task, or AX-CPT, an experimental paradigm designed to distill key aspects of cognitive control and context/goal maintenance (Braver and Cohen 1999; Braver, Barch, and Cohen 1999; Braver, Cohen, and Barch 2002). A key feature of this connectionist model, similar to the PBWM model discussed earlier by Frank and colleagues, is that contextual/goal representations are actively maintained in dorsolateral PFC, via mechanisms of recurrent connectivity and lateral inhibition. Most importantly, in this model, DA serves a joint neuromodulatory function within PFC, both gating representations into active maintenance (via phasic signals) and also regulating the persistence of maintenance

(via tonic signals; Braver, Barch, and Cohen 1999; Cohen, Braver, and Brown 2002). Model simulations with this DA neuromodulatory mechanism in PFC bolstered this hypothesis, providing evidence that context-dependent task performance, a key deficit in schizophrenia, is impaired with a noisy DA system (for more specific details, see Braver, Barch, and Cohen 1999). In particular, the model predicted very particular patterns of behavioral deficit in the AX-CPT task in participants with schizophrenia, as well as disruptions in the temporal dynamics of dorsolateral PFC activity, which were later confirmed experimentally (Barch et al. 2001; Braver, Cohen, and Barch 2002). Nevertheless, it has been difficult to demonstrate direct evidence that such deficits are specifically linked to DA neuromodulatory mechanisms, though recent advances in fMRI techniques have allowed researchers to more precisely measure dopaminergic phasic signals within the brainstem (D'Ardenne et al. 2012).

Evidence for a related account of contextual/goal representation deficits in schizophrenia was shown by Chambon et al. (2008). Here, the goal was to test the cascade model of hierarchical cognitive control in PFC proposed by Koechlin, Ody and Kouneiher (2003) to see whether it could account for particular patterns of behavioral impairment in individuals with schizophrenia. Interestingly, Chambon et al observed that sensory and episodic dimensions of cognitive control were preserved in patients with schizophrenia, whereas contextual control was impaired compared to matched healthy controls. In the study, patients generated significantly greater errors in tasks that required the ability to maintain context representations, and these impairments were highly correlated with disorganization score (e.g., a measure of disordered thought and behavior). Thus, the evidence is so far consistent with the hypothesis that in schizophrenia the ability to represent and actively maintain contextual or task–goal information is disrupted. In future investigations, it will be important to more directly test the claims of the cascade model that these deficits map appropriately along the rostro-caudal axis of PFC among individuals with schizophrenia.

4.4 Chapter Summary

This chapter highlighted several computational models that have played a seminal role in guiding theoretical accounts of cognitive control. We have selected these models because they provide promising testable hypotheses that have already stimulated a great deal of current experimental research, and which are likely to guide future investigations seeking to further elucidate the core neurocomputational mechanisms that underlie cognitive control. Furthermore, we hope that these models can be a useful primer for understanding computational approaches to cognitive processes more

broadly, as well as how these processes may be disrupted in mental illness. Although computational modeling approaches have played a central role in understanding normative cognitive function (e.g., memory, attention), many of these models have not yet been explicitly tested in psychiatric populations. Thus, we argue that developing accurate mechanistic models of normative cognitive functions can, in principle and in practice, facilitate greater insight into the etiology of psychopathology.

4.5 Further Study

Rumelhart et al. (1987) and O'Reilly & Munakata (2000) are seminal textbooks, both of which provide an in-depth introduction to connectionist computational models. The second book incorporates more biologically realistic algorithms and architectures, and explicitly accounts for extant cognitive neuroscience data. The authors have updated this work with more recent, free electronic editions, which include a chapter on executive function/cognitive control, available at https://grey.colorado.edu/CompCogNeuro/index.php/CCNBook/Book/.

For a review of the main scientific questions of cognitive control, and computational approaches that have been proposed to address these questions, see also O'Reilly et al. (2010) and Botvinick and Cohen (2014). An example of how different modeling levels can be utilized to provide converging evidence for cognitive control mechanisms can be found in Collins and Frank (2013). An example of how a computational model of cognitive control can be applied to make predictions about psychiatric disorder, specifically schizophrenia, is offered by Braver et al. (1999).