Special Series: Computational Psychiatry

ASSOCIATION FOR PSYCHOLOGICAL SCIENCE

Clinical Psychological Science
2015, Vol. 3(3) 378–399

Reprints and permissions:
sagepub.com/journalsPermissions.nav
DOI: 10.1177/2167702614565359
cpx.sagepub.com
SAGE

Model-Based Cognitive Neuroscience Approaches to Computational Psychiatry: Clustering and Classification

Thomas V. Wiecki, Jeffrey Poland, and Michael J. Frank
Department of Cognitive, Linguistic, and Psychological Science, Brown University

Abstract

Psychiatric research is in crisis. We highlight efforts to overcome current challenges by focusing on the emerging field of computational psychiatry, which might enable the field to move from a symptom-based description of mental illness to descriptors based on objective computational multidimensional functional variables. We survey recent efforts toward this goal and describe a set of methods that together form a toolbox to aid this research program. We identify four levels in computational psychiatry: (a) behavioral tasks that index various psychological processes, (b) computational models that identify the generative psychological processes, (c) parameter-estimation methods concerned with quantitatively fitting these models to subject behavior by focusing on hierarchical Bayesian estimation as a rich framework with many desirable properties, and (d) machine-learning clustering methods that identify clinically significant conditions and subgroups of individuals. As a proof of principle, we apply these methods to two different data sets. Finally, we highlight challenges for future research.

Keywords

computational psychiatry, drift diffusion model, computational cognitive neuroscience, RDoC, response inhibition

Received 9/27/13; Revision accepted 7/15/14

Imagine going to a doctor because of chest pain that has been bothering you for a couple of weeks. The doctor would sit down with you, listen carefully to your description of symptoms, and prescribe medication to lower blood pressure in case you have a heart condition. After a couple of weeks, your pain has not subsided. The doctor now prescribes medication against reflux, which finally seems to help. In this scenario, not a single medical analysis (e.g., electrocardiogram, blood work, or a gastroscopy) was performed, and medication with potentially severe side effects was prescribed on a trial-and-error basis. Although highly unlikely to occur if you walked into a primary care unit with these symptoms today, this scenario resembles much of contemporary psychiatry diagnosis and treatment.

There are several reasons for this discrepancy in sophistication between psychiatry and other fields of medicine. First and foremost, mental illness affects the brain—the most complex biological system yet encountered. Compared with the level of scientific understanding achieved on other organs of the human body, such as the heart, our understanding of the normally functioning brain is still, arguably, in its infancy.

Despite this complexity, concerted efforts in the brain sciences have led to an explosion of knowledge and understanding about the healthy and diseased brain in the past decades. The discovery of highly effective psychoactive drugs in the 1950s and 1960s raised expectations that psychiatry would progress in a similar fashion. It is unfortunate that, in retrospect, it appears that these discoveries were serendipitous in nature, given that little progress has been made since (e.g., Hyman, 2012; Insel et al., 2010). This lack of progress also has caused many major pharmaceuticals companies, such as AstraZeneca and GlaxoSmithKline, to withdraw from psychiatric drug development and to close large research centers (Cressey,

Corresponding Authors:
Michael J. Frank and Thomas V. Wiecki, Department of Cognitive, Linguistic, and Psychological Science, Brown University, 190 Thayer St., Providence, RI 02912-1821
E-mail: michael_frank@brown.edu and thomas.wiecki@gmail.com

2011; Nutt & Goodwin, 2011). In addition, research on mental illness, based on conventional psychiatric diagnostic categories and practices (as reflected in the various editions of the *Diagnostic and Statistical Manual of Mental Disorders, DSM*; e.g., American Psychiatric Association, 2013), has been widely viewed as disappointing, and the *DSM* system of classification itself has been viewed as an impediment to more productive research. As a consequence, psychiatry is a field in crisis (Hyman, 2012; Insel et al., 2010; Poland, Von Eckardt, & Spaulding, 1994; Sahakian, Malloch, & Kennard, 2010). As outlined in more detail later, a central issue is a lack of sufficiently powerful theoretical and methodological resources for managing the features of mental illness (e.g., a lack of measurable quantitative descriptors). This lacuna prevents effective management of the multidimensional hierarchical complexity, dynamic interactivity, causal ambiguity, and heterogeneity of mental illness. And it leads to an explanatory gap of how basic neurobiological processes and other causes result in complex disorders of the mind (Hyman, 2012; Montague, Dolan, Friston, & Dayan, 2011).

In the present study, we review current challenges in psychiatry and recent efforts to overcome them. Several examples from the domain of decision making show the promise of moving away from symptom-based description of mental illness and instead formulating objective, quantifiable computational biomarkers as a basis for further psychiatric research. We then introduce a computational cognitive toolbox that is suited to construct these computational biomarkers. We focus on sequential sampling models (SSMs) of decision making, which serve as a case study for how computational models, when fit to behavior, have successfully been used to identify and quantify latent neurocognitive processes in healthy humans. Bayesian methods provide a resourceful framework to fit these models to behavior and establish individualized descriptors of neurocognitive function. After establishing the validity of these models to provide neurocognitive descriptors of individuals, we review how clustering techniques can be used to construct a map of individual differences based on these neurocognitive descriptors.

To demonstrate the viability and potential of these methods, we reanalyze two data sets, thereby providing a proof of principle before discussing future challenges in application to psychiatric populations. The first data set consists of a group of young and old subjects who performed three different decision-making tasks (Ratcliff, Thapar, & McKoon, 2010). After fitting subjects' choices and response-time (RT) distributions with the drift-diffusion model (DDM) using hierarchical Bayesian parameter estimation, we provide each subject's parameter estimates as inputs to an unsupervised clustering algorithm. We show that the clustering is sensitive to age after nuisance variables are regressed out and that this clustering shows consistently better recovery of the age-groups than if behavioral summary statistics (e.g., mean RT and accuracy) are used alone. Moreover, factor analysis on the computational parameters extracts meaningful latent variables that describe cognitive ability. For this data set, no identified brain-based mechanism was analyzed. In contrast, for the second data set, we relied on a hypothesis-driven approach that suggested a mechanism for how a specific decision parameter—the decision threshold—varies as a function of activity communicated between frontal cortex and the subthalamic nucleus (STN). A previous study showed that STN deep-brain stimulation disrupted decision-threshold regulation across a group of patients with Parkinson's disease (PD; Cavanagh et al., 2011). In the following, we show that we can classify individual patients' brain-stimulation status (off or on) with relatively high accuracy, given model parameters, and better than that achieved on the basis of brain-behavior correlations alone.

Current Challenges in Psychiatry

The current crisis in psychiatry has complex causes that are deeply rooted in existing classification systems (e.g., *DSM*, International Classification of Diseases). In this section, we identify some of the problems these systems introduce and provide indications of the sorts of resources required for more productive research programs. In the subsequent section, we review recent attempts to meet these challenges and the sorts of resources that have been introduced for this purpose. As other researchers before us have done, we proceed to suggest an approach to research of mental disorders that aims to link cognitive and pure neuroscience to mental illness without the restrictions of prior classification schemes (Robbins, Gillan, Smith, de Wit, & Ersche, 2012).

DSM *and research*

For decades, the *DSM* has been the basis of clinical diagnosis, treatment, and research of mental illness. At its core, the *DSM* defines distinct disorder categories, such as schizophrenia (SZ) and depression, in a way that is atheoretical (i.e., with no reference to specific causal hypotheses) and focused on clinical phenomenology. Thus, these categories are mainly derived from translating subjective experience to objective symptomatology while assuming unspecified biological, psychological, or behavioral dysfunctions (Poland et al., 1994).

Although primarily intended to be of value to clinicians, the *DSM* has also played a substantial role as a classification system for scientific research with the goals of validating the diagnostic categories and translating research results directly into clinical practice. Although these research goals are commendable, decisions regarding systematic classification are more often based on perceptions of clinical utility rather than scientific merit. As a consequence, *DSM*-based research programs have failed to deliver consistent, replicable, and specific results, and it has been widely observed that the validation of *DSM* categories has been limited, that *DSM* categories do not provide well-defined phenotypes, and that they have limited research utility.

Heterogeneity and comorbidity

One major problem of contemporary psychiatric classification is the heterogeneity of individuals receiving identical diagnoses. With respect to symptomatology, one striking example of this is SZ, with regard to which one must exhibit at least two out of five symptoms to receive a diagnosis. It is thus possible to have patients with completely different symptomatology being diagnosed as schizophrenic. It is important, however, that problems of heterogeneity concern more than just symptoms; there is probably heterogeneity at all levels of analysis, including heterogeneity of causal processes (Poland et al., 1994). And, as we discuss later, such heterogeneity is not just a feature of clinical populations but also may be a feature of the general population. As a consequence, heterogeneity poses a serious challenge for research (e.g., it introduces uncontrolled sources of variance, it limits the generalizability of results) and points to the necessity of developing techniques for its management.

Comorbidity is widely believed to constitute a second major problem for psychiatric classification. Defined as the co-occurrence of multiple disorders in one individual, it has been widely documented that “comorbidity between mental disorders is the rule rather than the exception, invading nearly all canonical diagnostic boundaries” (Buckholtz & Meyer-Lindenberg, 2012, p. 996). It is important to differentiate between two relevant types of comorbidity: True comorbidity is a result of independent disorders co-occurring; artificial comorbidity is a result of separately classifying disorders that have overlapping symptom criteria, have a common cause, or share a pathogenic cascade. This distinction points to a more general problem concerning the management of causal ambiguity that is found at the level of symptoms but also at other levels of analysis. Specifically, the problem is one of identifying which causal structures and processes produce a given clinical presentation or a given pattern of functioning at some other level; because clinical presentations and patterns of functioning can be produced by different causal structures and processes, the challenge for researchers is to develop techniques for identifying and managing such causal ambiguity.

In addition to challenges of heterogeneity and comorbidity, several other features of the domain of mental illness pose challenges to research and require sophisticated tools and techniques for their effective management. These include hierarchical organization of the brain and various sorts of interlevel relationship and coordination (e.g., the explanatory gap), dynamic interactivity, multidimensional complexity, context sensitivity, identification of norms of functioning, and identification of meaningful groupings of individuals. As we discuss herein, each of these features creates problems that contribute to an understanding of why the current crisis in research exists and of the sorts of resources and strategies required for more productive research programs.

Potential Solutions

As summarized in the preceding discussion, the shortcomings of the current *DSM* classification system and the problems they pose for research are well documented. In the following, we outline some current efforts to address these challenges.

Research Domain Criteria Project and a Roadmap for Mental Health Research in Europe

The Research Domain Criteria Project (RDoC) is an initiative by the National Institute of Mental Health (Insel et al., 2010). RDoC improves on previous research efforts based on the *DSM* in the following ways. First, as the name implies, it is conceptualized as a research framework only and, thus, is clearly separated from clinical practice. Second, RDoC is completely agnostic about *DSM* categories. Instead of a top-down approach that aims to identify neural correlates of psychiatric disorders, RDoC suggests a bottom-up approach that builds on the current understanding of neurobiological underpinnings of different cognitive processes and links those to clinical phenomena. Third, the RDoC research program integrates data from different levels of analysis, such as imaging, behavior, and self-reports.

At its core, RDoC is structured into a matrix with columns representing different “units of analysis” and rows for research domains. The units of analysis include genes, molecules, cells, circuits, physiology, behavior, and self-reports. Research domains are clustered into negative- and positive-valence systems, cognitive systems, systems for social processes, and arousal/regulatory systems. Each of these domains is further subdivided into distinct processes; for example, cognitive systems include attention, perception, working memory, declarative memory, language behavior, and executive control.

Despite clear improvements over previous *DSM*-based research programs, the RDoC initiative currently lacks explicit consideration of computational descriptors. As outlined later, computational methods show great promise to help link different levels of analysis, elucidate clinical symptoms, and identify subgroups of healthy control (HC) and patient populations.

More recent, the European Commission started the Roadmap for Mental Health Research in Europe (ROAMER) initiative with the goal of better integrating biomedicine, psychology, and public-health insights to further research into mental illnesses.

Neurocognitive phenotyping

In a recent review article, Robbins et al. (2012) suggested the use of neurocognitive endophenotypes to study mental illness: "Neurocognitive endophenotypes would furnish more quantitative measures of deficits by avoiding the exclusive use of clinical rating scales, and thereby provide more accurate descriptions of phenotypes for psychiatric genetics or for assessing the efficacy of novel treatments" (p. 82).

Of particular interest are three studies that use such neurocognitive endophenotypes by constructing multidimensional profiles (MPs) from behavioral summary statistics across a battery of various neuropsychological tasks used to identify subtypes of attention-deficit/hyperactivity disorder (ADHD; Durston et al., 2008; Fair, Bathula, Nikolas, & Nigg, 2012; Sonuga-Barke, 2005).

Durston et al. (2008) argued that there are distinct pathogenic cascades within at least three different brain circuits that can lead to symptomatology involved in ADHD. Specifically, abnormalities in dorsal frontostriatal, orbito-frontostriatal, or fronto-cerebellar circuits can lead to impairments of cognitive control, reward processing, and timing, respectively. Core deficits in one or multiple of these brain networks can thus result in a clinical diagnosis of ADHD and provide a compelling explanation for the heterogeneity of the ADHD patient population. Preliminary evidence for this hypothesis is provided by Sonuga-Barke (2005), who used principal component analysis on MPs (based on a neuropsychological task battery) of ADHD patients and identified three distinct subtypes that covaried on timing, cognitive control, and reward.

A similar approach to identifying clusters in the ADHD population using MPs was taken by Fair et al. (2012). The authors applied graph theory to identify individual behavioral functional clusters within not only the ADHD patient population but also HC subjects. It is interesting that the authors found that HC and ADHD is not the predominant dimension along which clusters form. Instead, Fair et al. uncovered different functional profiles (e.g., one cluster might show differences in response inhibition, whereas another shows differences in RT variability), each of which contained both HC and patient subgroups. Nevertheless, and critically, a classifier trained to predict diagnostic category achieved better performance when classifying within each functional profile than did a classifier trained on the aggregated data. In other words, this implies that the overall population clusters into different cognitive profiles, and ADHD affects individuals differently on the basis of which cognitive profile they exhibit. The results of this study suggested that the source of heterogeneity not only may be distinct pathogenic cascades being labeled as the same disorder but also may be a result of the inherent heterogeneity present in the overall population—healthy and disordered.

The studies discussed all exemplify the danger of lumping subjects at the level of symptoms and treating them as one homogeneous category with a single, identifiable pathological cascade. Instead, these studies used MPs to find an alternative characterization of subjects independent of their *DSM* classification that is (a) quantitatively measurable, (b) a closer approximation to the underlying neurocircuitry (Robbins et al., 2012), and (c) cognizant of heterogeneity in the general population.

Nevertheless, this approach still has problems. First, although there was less reliance on *DSM* categories, these studies still used the diagnostic label for recruiting subjects, selecting tasks, framing and testing hypotheses, and drawing inferences. It could be imagined, for example, that patients with compulsive disorders, such as obsessive-compulsive disorder (OCD) or Tourette's syndrome, have abnormalities in similar brain circuits and, consequently, pathologies, deficits, and impairments may crosscut these (and other) diagnostic categories. Thus, if only ADHD patients are recruited, a critical part of the picture might be missed. Second, the cognitive-task battery covers only certain aspects of cognitive function. Other tasks that, for example, measure working memory or reinforcement learning (RL), both of which involve frontostriatal function, would be a useful addition to help resolve causal ambiguity. More specifically, performance on each individual task is assessed by an aggregate performance score. Recent behavioral and neuropsychological findings, however, have suggested that executive control (as an example) in a single task may instead be more accurately characterized as a collection of related but separable abilities, a pattern referred to as the unity and diversity of executive functions. Furthermore, most cognitive tasks rely on a concerted and often intricate interaction of various neural networks and cognitive processes (see, e.g., Collins & Frank, 2012). This task-impurity problem complicates identification of separate functional impairments and brain circuits solely on the basis of MPs.

In sum, although cognitive phenotypes provide a useful framework for measuring brain function, there is still ambiguity if behavioral scores that provide an aggregate measure of various brain networks are used. The idea that a neural circuit can contribute to different cognitive functions helps explain why diverse mental illnesses can exhibit similar symptoms (comorbidity; Buckholtz & Meyer-Lindenberg, 2012). Disentangling these transdiagnostic patterns of psychiatric symptoms thus requires identification and measurement of underlying brain circuits and functions. Whereas Buckholtz and Meyer-Lindenberg (2012) proposed the use of functional imaging studies and genetic analysis, we discuss how computational modeling can contribute to disambiguate the multiple pathways leading to behavioral features.

Computational psychiatry

Computational models at different levels of abstraction have had tremendous impact on the field of cognitive neuroscience. The aim is to construct models based on integrated evidence from neuroscience and psychology to explain neural activity as well as cognitive processes and behavior. Although more detailed biologically inspired models, such as biophysical and neural-network models, are generally more constrained by neurobiology, they often have many parameters that make them less suitable to fit them directly to human behavior. Conversely, more abstract, algorithmic models often have fewer parameters that allow them to be fit directly to data at the cost of being less detailed about the neurobiology. Normal linking of one level of analysis to another is useful to identify plausible neural mechanisms that can be tested with quantitative tools. Critically, all of these models allow for increased specificity in the identification of different neuronal and psychological processes that are often lumped together in analyses of task behavior based on summary statistics.

The nascent field of computational psychiatry uses computational models to infer dysfunctional latent processes in the brain. Montague et al. (2011) defined the goal for computational psychiatry as

> extract[ing] normative computational accounts of healthy and pathological cognition useful for building predictive models of individuals. . . . Achieving this goal will require new types of phenotyping approaches, in which computational parameters are estimated (neurally and behaviorally) from human subjects and used to inform the models. (p. 75)

More generally, the tools and techniques of computational cognitive neuroscience (e.g., modeling at multiple levels of analysis, parameter estimation, classification algorithms) are especially well suited for representing and managing the various features of mental illness identified earlier (e.g., hierarchical and multidimensional organization, nonlinear dynamic interactivity, context sensitivity, heterogeneity, and individual variation). Thus, computational psychiatry holds out considerable promise as a research program directed at mental illness.

On the basis of this approach, Maia and Frank (2011) identified computational models as a

> valuable tool in taming [the complex pathological cascades of mental illness] as they foster a mechanistic understanding that can span multiple levels of analysis and can explain how changes to one component of the system (for example, increases in striatal D2 receptor density) can produce systems-level changes that translate to changes in behavior. (p. 154)

Moreover, three concrete strategies for how computational models can be used to study brain dysfunction were defined:

- Deductive approach: Established neuronal or neural-circuit models can be tested for how pathophysiologically plausible alterations in neuronal state, for instance, connectivity or neurotransmitter levels (e.g., dopamine is known to be reduced in PD), affect system-level activations and behavior. This is essentially a bottom-up approach, given that it involves the study of how known or hypothesized neuronal changes affect higher-level functioning.
- Abductive approach: Computational models can be used to infer neurobiological causes from known behavioral differences. In essence, this is a top-down approach that tries to link behavioral consequences back to underlying latent causes.
- Quantitative abductive approach: Parameters of a computational model are fit to a subject's behavior on a suitable task or task battery. Different parameter values point to differences in underlying neurocircuitry of the associated subject or subject group. These parameters can be used either comparatively to study group differences (e.g., healthy and diseased) or as a regressor with, for example, symptom severity. This approach is more common with abstract models than with neural-network models, given that the former typically have fewer parameters and, thus, can be more easily fit to data.

Case studies in the domain of decision making

One key area in which computational models have had tremendous success is in the elucidation of how the different cognitive and neurobiological gears work together in the domain of decision making. Many mental illnesses can be characterized by aberrant decision making of one sort or another (Maia & Frank, 2011; Montague et al., 2011). In the following section, we review recent cases in which computational models of decision making have been used to better understand brain disorders.

Computational models of RL: PD and SZ. Our first case study concerns PD. Its most visible symptoms affect the motor system as manifest in hypokinesia, bradykinesia, akinesia, rigidity, tremor, and progressive motor degeneration. However, cognitive symptoms recently have received increased attention. PD is an intriguing neuropsychiatric disorder because its core pathology is well identified to be the cell death of midbrain dopaminergic neurons in the substantia nigra pars compacta. Neural-network models of the basal ganglia interpret this brain network as an adaptive action-selection device that conditionally gates internal or external actions on the basis of their previous reward history, which is learned via dopaminergic signals (Montague, Dayan, & Sejnowski, 1996; Schultz, Dayan, & Montague, 1997). Behavioral RL tasks show that the chronic low levels of dopamine in PD patients result in a bias toward learning from negative reward-prediction errors (RPEs) at the cost of learning from positive RPEs. In extension, we have argued that PD is not a motor disorder per se but rather an action-selection disorder in which the progressive decline of motor and cognitive function can be interpreted in terms of aberrant learning and not to select actions.

In this case study, an existing biological model of healthy brain function was paired with a known and well-localized neuronal dysfunction to extend our understanding of the symptomatology of a brain disorder and to reconceive the nature of the dysfunctions involved. Note, however, that the model was not fit to data quantitatively, nor were MPs provided to resolve residual causal ambiguity associated with the task-impurity problem. In the terminology established by Maia and Frank (2011), this is an example of the *deductive* approach, in which the model provides a mechanistic bridge that explains how abnormal behavior can result from neurocircuit dysfunctions.

Despite SZ being the focus of intense research during the past decades, no single theory of its underlying neural causes has been able to explain the diverse set of symptoms that lead to an SZ diagnosis. Current psychiatric practices view the symptomatology of SZ as structured in terms of positive symptoms, such as psychosis; negative symptoms, such as anhedonia, which refers to the inability to experience pleasure from activities usually found enjoyable, such as social interaction; and cognitive deficits (Elvevåg & Goldberg, 2000).

Recent progress has been made by the application of RL models to understand individual symptoms or a single symptom category (e.g., negative symptoms) rather than SZ as a whole (J. M. Gold et al., 2012; J. M. Gold, Waltz, Prentice, Morris, & Heerey, 2008; Strauss et al., 2011; Waltz, Frank, Wiecki, & Gold, 2011).

Using an RL task, Waltz, Frank, Robinson, and Gold (2007) found that SZ patients showed reduced performance in selecting previously rewarded stimuli compared with HC subjects and that this performance deficit was most pronounced in patients with severe negative symptoms. It is notable that patients with SZ and HC subjects did not differ in their ability to avoid actions leading to negative outcomes. However, as a result of the task-impurity problem, this behavioral analysis did not allow researchers to differentiate whether SZ patients were impaired at learning from positive outcomes or from a failure in representation of the prospective reward values during decision making. The following is a strategy for resolving this problem.

This dichotomy in learning versus representation is also present in two types of RL models—actor-critic and Q-learning models (Sutton & Barto, 1998). An actor-critic model consists of two modules: an actor and a critic. The critic learns the expected rewards of states and trains the actor to perform actions that lead to better-than-expected outcomes. The actor itself learns only “action propensities,” in essence, stimulus-response links. Q-learning models, conversely, learn to associate actions with their reward values in each state. Thus, whereas a Q-learning model has an explicit representation of which action is most valued in each state, the actor-critic model will choose actions on the basis of those that have previously yielded positive prediction errors—regardless of whether those arose from an unexpected reward or the absence of an expected loss. Thus, the differences between these two models can be exploited to attempt to resolve the causal ambiguity exhibited by the results discussed.

In a follow-up study, J. M. Gold et al. (2012) administered a new task that paired a neutral stimulus in one context with a positive stimulus and in another context with a negative stimulus. Although the neutral stimulus has the same value of zero in both contexts, it is known that dopamine signals RPEs that drive learning in the basal ganglia and code outcomes *relative* to the expected reward (Montague et al., 1996; Schultz et al., 1997). Thus, in the negative context, receiving nothing is better than expected and will result in a positive RPE, thereby driving learning in the basal ganglia to select this action in the future (Maia, 2010). In a test period in which no rewards were presented, subjects had to choose between an action that had been rewarding and one that had avoided a loss. Both actions should have been associated with better-than-expected outcomes. An actor-critic model should thus show a tendency to select the neutral stimulus, whereas a Q-learning model with representation of the reward contingencies should mainly select the one with a higher reward. It is intriguing that when both of these models were fit to subject data, the actor-critic model produced a better fit for SZ patients with a high degree of negative symptoms, whereas HC subjects and SZ patients with low negative symptoms were better fit by a Q-learning model. In other words, patients with negative symptoms largely based decisions on learned stimulus-response associations instead of expected reward values. It is notable that HC subjects and the low-negative-symptom SZ group did not differ significantly in their RL behavior. This study by J. M. Gold et al. demonstrated how computational analyses can differentiate between alternative mechanisms that can explain deficiencies in reward-based choice. Many RL tasks can be solved by learning either stimulus-response contingencies or expected reward values (or both), but the model and appropriate task manipulation allows one to extract to which degree these processes are operative and, thus, helps to resolve the task-impurity problem.

In a related line of work, Strauss et al. (2011) tested HC subjects and SZ patients on an RL task that allowed subjects to either adopt a safe strategy and *exploit* the rewards of actions with previously experienced rewards or *explore* new actions with perhaps even higher payoffs. Frank, Doll, Oas-Terpstra, and Moreno (2009) developed a computational model that can recover how individual subjects balance this exploration-exploitation trade-off. It is intriguing that in applying this model to SZ patients, Strauss et al. found that patients with high anhedonia ratings were less willing to explore their environment and uncover potentially better actions. This result suggests a reinterpretation of the computational cognitive process underlying lack of social engagement associated with anhedonia. For example, one might assume that the lack of engagement of social activities of anhedonistic patients results from an inability to experience pleasure and, as a consequence, a failure to learn the positive value of social interaction. Instead, this study suggested that lack of social engagement associated with anhedonia is a result of an inability to consider the prospective benefit of doing something that might lead to better outcomes. These results also lead to the prediction that patients with SZ would not, for example, seek out new social interactions (because of the low value placed on exploration) but could still enjoy social interactions once established. Again, computational strategies allow for a reconceptualization and disambiguation of clinical phenomena.

In sum, J. M. Gold et al. (2012) and Strauss et al. (2011) used a *quantitative abductive* approach to infer aberrant computational cognitive processes in RL in a subgroup of SZ patients. By grouping subjects according to symptom type and severity instead of diagnosis, the authors identified more refined research targets and addressed the problem of heterogeneity. By combining models and strategically designing task demands, J. M. Gold et al. pursued an innovative strategy for resolving problems of interpretation resulting from task impurity.

Another relevant line of work includes that of Brodersen et al. (2013), who used dynamic causal modeling (Friston, Harrison, & Penny, 2003)—a Bayesian framework for inferring network connectivity between brain areas from functional MRI (fMRI) data—on HC subjects and SZ patients performing a numerical *n*-back working memory task. Supervised learning methods demonstrated a clear benefit (71% accuracy) of using dynamic causal modeling compared with more traditional methods, such as functional connectivity (62%). Moreover, clustering methods were sensitive to various SZ subtypes, which showed the potential of this approach to identify clinically meaningful groups in an unsupervised manner. Finally, we refer to Huys et al. (2012) for an example of how a computational-psychiatry analysis can be used to relate depressive-symptom severity to a specific cognitive process involved in planning multiple future actions.

Computational models of response inhibition. Besides RL, response inhibition is another widely studied phenomenon in cognitive neuroscience relevant to mental illness. Response inhibition is required when actions in the planning or execution stage are no longer appropriate and must be suppressed. The antisaccade task is one such task that is often used in a psychiatric setting (e.g., Aichert et al., 2012; Fukumoto-Motoshita et al., 2009). It requires subjects to inhibit a prepotent response to a salient stimulus and instead saccade to the opposite side (Hallett, 1978). A wealth of literature has demonstrated reduced performance of psychiatric patients with disorders, including ADHD (Nigg, 2001; Oosterlaan, Logan, & Sergeant, 1998; Schachar & Logan, 1990), OCD (Chamberlain, Fineberg, Blackwell, Robbins, & Sahakian, 2006; Menzies et al., 2007; Morein-Zamir, Fineberg, Robbins, & Sahakian, 2010; Penadés et al., 2007), SZ (Badcock, Michie, Johnson, & Combrinck, 2002; Bellgrove et al., 2006; Huddy et al., 2009), PD (van Koningsbruggen, Pender, Machado, & Rafal, 2009), and substance-abuse disorders (Monterosso, Aron, Cordova, Xu, & London, 2005; Nigg et al., 2006). However, as demonstrated by Wiecki and Frank (2013), even a supposedly simple behavioral task, such as the antisaccade task, requires a finely orchestrated interplay between various brain regions, including frontal cortex and basal ganglia. It thus cannot be said that decreased accuracy in this task is evidence of response inhibition deficits per se, given that the source of this performance impairment can be manifold (i.e., the antisaccade task exhibits the task-impurity problem).

In sum, the use of computational models that allow mapping of behavior to psychological processes could thus be categorized as the *computational abductive* approach. However, in addition to managing the task-impurity problem just mentioned, ambiguity of how psychological processes relate to the underlying neurocircuitry still has to be resolved. By combining different levels of modeling, researchers can better identify and study these ambiguities. Ultimately, this might allow development of tasks that use specific conditions (e.g., speed-accuracy trade-off, reward modulations, and conflict) to disambiguate the mapping of psychological processes to their neurocircuitry. The use of biological-process models to test different hypotheses about the behavioral and cognitive effects of neurocircuit modulations would correspond to the *deductive* approach. In other words, by combining the research approaches outlined by Maia and Frank (2011), we can use our understanding of the different levels of processing to inform and validate how these levels interact in the healthy and dysfunctional brain.

Thus, there are a few example studies in which researchers have applied established computational models to identify model parameters (which aim to describe specific cognitive functions) and related them to the severity of a specific clinical symptom or used them to identify measureable cognitive impairments. Such targets (viz., specific symptoms, measureable impairments) represent more refined research targets than do *DSM* diagnostic categories. In addition, through the use of strategically designed task batteries and MPs, problems of heterogeneity and task impurity can be managed. And the combination of various research approaches (e.g., multiple modeling strategies, task batteries and MPs, task manipulations, novel approaches to sampling) can provide a strategic framework for studying relations between neural and computational levels of analysis in mental illness.

Levels of Computational Psychiatry

Thus far, we have identified a variety of challenges to research concerning mental illness and various strategies that have been employed to meet those challenges. Special attention has been given to computational psychiatry as an especially promising research program. In all cases, promise for effectively meeting the research challenges depends on the availability of conceptual and representational resources and associated strategies and techniques that are sufficiently powerful, given the features of the domain of mental illness and the problems it poses for research.

In this section, we provide an overview of a four-level approach to the computational analysis of cognitive function and dysfunction by focusing on decision making and using SSMs as a concrete example (see Table 1 for a delineation of terminology applicable to our discussion). Such models provide a versatile tool to model cognitive function, but fitting such models to data presents significant technical challenges as well. In the following, we identify four levels of the analysis: Level 1, strategic identification of cognitive tasks to be employed for the collection of performance data; Level 2, the fitting of computational models to the performance data; Level 3, parameter estimation; and Level 4, identification of clusters and relations to clinical symptom severity (see Fig. 1 for an overview). We show how hierarchical Bayesian modeling and Bayesian mixture models can be deployed to engage a variety of challenges at the various levels of the analysis. Subsequently, we demonstrate the use of these methods on two data sets as a "proof of concept." The methods identified in this section have direct applicability to the analysis of cognitive functions in mental illness.

Level 1: Cognitive tasks

Cognition spans many mental processes that include attention, social cognition, memory, emotion, decision making, and reasoning, to name a few. Various subfields devoted to each of these have developed a range of cognitive tasks that purport to reveal the underlying mechanisms. Research in computational psychiatry can draw on these tasks to create task batteries for the collection of performance data usable for the analysis of cognitive function; both the sensitivity and the specificity of tasks to cognitive functions are important characteristics, although the task-impurity problem complicates the analysis of data and their use in isolating and specifying cognitive functions. Rather than provide a list of tasks used (see the case studies discussed earlier for some examples), we discuss desirable properties that cognitive tasks should exhibit. A single cognitive task used in computational psychiatry ideally should be tuned to assess a specific cognitive function, separable from others; this is enabled by the following:

- a task analysis that identifies what functions are engaged and how they are engaged;
- parsimony in relying on as few cognitive processes as possible;
- stress on cognitive processing in some way to reveal break-off points and allow a sensitive measure of the target function;
- an established theory regarding the neural correlates of the target functions; and
- an established computational model that links behavior to psychological-process parameters.

Given the task-impurity problem and other forms of causal ambiguity, task batteries ideally should be strategically constructed to measure a range of relevant cognitive functions and other variables to aid in the interpretation of task performance and the isolation of specific functions and dysfunctions. This can be achieved by including covarying

Table 1. Terminology

Term	Definition
Psychological-process model	A computational model that tries to parameterize the cognitive processes underlying behavior. This class of models is not primarily concerned with neural implementations of these processes. Often these models have a parsimonious parameterization that allows them to be fit to behavior.
Drift-diffusion model	An evidence-accumulation model used in decision-making research.
Reinforcement learning	Learning to adapt behavior to maximize rewards and minimize punishment.
Parameter estimation/fitting	The process of finding parameters that best capture the behavior on a certain task.
Bayesian modeling	A parameter-estimation method that allows for great flexibility in defining structure and prior information about a certain domain.
Comorbidity	The co-occurrence of multiple disorders in one individual.
Heterogeneity	The fact that there is systematic variation between subjects diagnosed with the same mental illness.
Task-impurity problem	The fact that no single cognitive task measures just one construct but that task performance is a mixture of distinct cognitive processes.
Multidimensional profile	A multidimensional descriptor of a subject's cognitive abilities as measured by summary statistics (e.g., accuracy) of cognitive tasks spanning multiple cognitive domains.
Computational multidimensional profile	A multidimensional profile that includes parameters estimated from a psychological-process model that (a) more directly relates to cognitive ability and (b) deconstructs different cognitive processes that contribute to individual task performance (i.e., task-impurity problem).

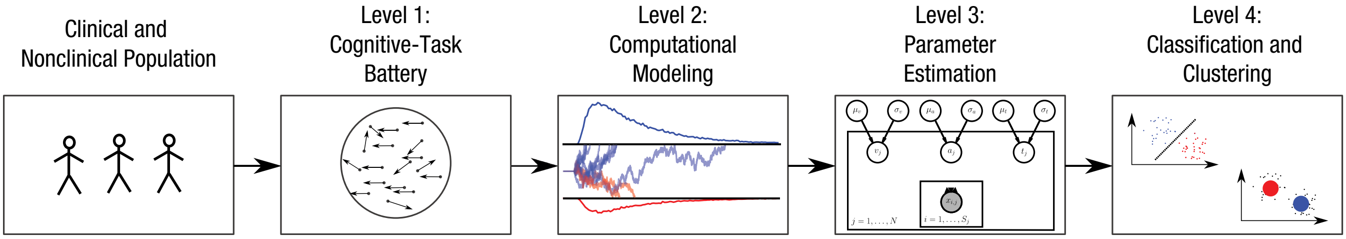

Fig. 1. Illustration of the four levels of computational psychiatry. Clinical and nonclinical populations are tested on a battery of cognitive tasks. Computational models can relate raw task performance (e.g., response time and accuracy) to psychological and neurocognitive processes. These models can be estimated via various methods (depicted is a simplified graphic of the HDDM or hierarchical drift-diffusion model). Finally, on the basis of a resulting computational multidimensional profile, supervised and unsupervised learning algorithms can be trained to either predict disease state, uncover groups and subgroups in clinical and healthy populations, or relate model parameters to clinical symptom severity.

factors (i.e., conditions) in individual tasks that affect only one mental function, which can then be identified. For example, Collins and Frank (2012) were able to separately estimate the contributions of working memory and RL in a single task by testing multiple conditions that increased load on working memory alone. Because working memory contributions can contaminate the estimation of the RL component, this manipulation enabled a model to not only capture the WM component but also better estimate the RL component.

Level 2: Computational models

Computational models in cognitive neuroscience exist on various levels of abstraction that range from biophysical neuronal models to abstract psychological-process models. Although each of these is informative in its own regard in elucidating mental function and dysfunction, we focus here on psychological-process models. This class of model has the unique advantage of being simple enough so that it can be fit directly to behavior; that is, it is preferred, from a statistical analysis point of view, given the level of data collected. The fitted parameters often quantify cognitive ability in terms of psychological-process variables rather than behavioral summary statistics. For example, in a simple detection task, one might consider the RT speed as a good measure of task performance. However, by adjusting the speed-accuracy trade-off, mean RT can easily be shortened just by increasing the false alarm rate. This obviously would not indicate an individual's superior processing abilities. An SSM analysis, however, would be able to disentangle response caution (i.e.,

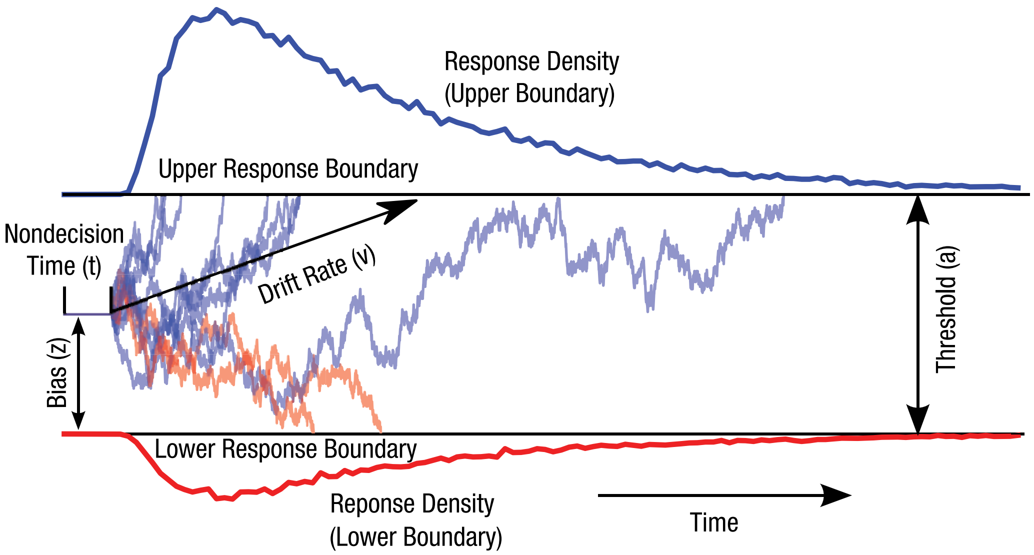

Fig. 2. Trajectories of multiple drift processes (blue and red lines, middle panel). Evidence is accumulated over time (*x*-axis) with drift rate (v) until one of two boundaries, separated by threshold (a), is crossed and a response is initiated. Upper (blue) and lower (red) panels contain histograms over boundary-crossing times for two possible responses. The histogram shapes match closely to that observed in response-time measurements of research subjects.

decision threshold) and processing abilities (i.e., drift rate): These are generative parameters that produce the joint distribution of accuracy and RT. Intuitively, an increase in decision threshold would lead to more accurate but slower responses, whereas an increase in drift rate would lead to higher accuracy but also faster responses (Ratcliff & McKoon, 2008). In the following section, we present a simulation experiment that shows how two groups can be clearly separated in their DDM parameters but strongly overlap when described in terms of RT and accuracy summary statistics.

SSMs. As outlined earlier, RL models have already proven to be a valuable tool in explaining neuropsychological disorders and their symptoms. A computational-psychiatric framework that aims to explain the multifaceted domain of mental illness must include computational cognitive neuroscience models that cover a broad range of cognitive processes (see, e.g., O'Reilly, Munakata, Frank, Hazy, & Contributors, 2012, for a broad coverage of such models). We focus on SSMs as an illustrative example of how these models have been applied to study normal and aberrant neurocognitive phenomena, how they can be fit to data using Bayesian estimation, and how subgroups of similar subjects can be inferred using mixture models.

SSMs (e.g., Townsend & Ashby, 1983), such as the DDM, have established themselves as the de facto standard for modeling data from simple decision-making tasks (e.g., Smith & Ratcliff, 2004). Each decision is modeled as a sequential extraction and accumulation of information from the environment or internal representations. Once the accumulated evidence crosses a threshold, a corresponding response is executed. This simple assumption about the underlying psychological process has the important property of reproducing not only choice probability and mean RT but also the entire distribution of RTs separately for accurate and erroneous choices in simple two-choice decision-making tasks. It is interesting that this evolution of the decision signal in SSMs can also be interpreted as a Bayesian update process (e.g., Bitzer, Park, Blankenburg, & Kiebel, 2014; Deneve, 2008; J. I. Gold & Shadlen, 2002; Huang & Rao, 2013). This may be useful because it would place SSMs under a more axiomatic framework and prevent the impression that SSMs are merely convenient heuristics.

The DDM models decision making in two-choice tasks. Each choice is represented as an upper and lower boundary. A drift process accumulates evidence over time until it crosses one of the two boundaries and initiates the corresponding response (Ratcliff & Rouder, 1998; Smith & Ratcliff, 2004). The speed with which the accumulation process approaches one of the two boundaries is called the drift rate and represents the relative evidence for or against a particular response. Because there is noise in the drift process, the time of the boundary crossing and the selected response will vary between trials. The distance between the two boundaries (i.e., threshold) influences how much evidence must be accumulated until a response is executed. A lower threshold makes responding faster in general but increases the influence of noise on decision making, whereas a higher threshold leads to more cautious responding. RT, however, is not solely composed of the decision-making process—perception, movement initiation, and execution all take time and are summarized into one variable called nondecision time. The starting point of the drift process relative to the two boundaries can influence whether one response has a prepotent bias. This pattern gives rise to the RT distributions of both choices (see Fig. 2 for trajectories of multiple drift processes; mathematical details of the methods motivated herein can be found in the Drift-Diffusion Model section in the Supplemental Material available online).

Relationship to cognitive neuroscience. SSMs were originally developed from a pure information-processing point of view and primarily used in psychology as a high-level approximation of the decision process. More recent efforts in cognitive neuroscience have simultaneously (a) validated core assumptions of the model by showing that neurons indeed integrate evidence probabilistically during decision making (J. I. Gold & Shadlen, 2007; Smith & Ratcliff, 2004) and (b) applied this model to describe and understand neural correlates of cognitive processes (e.g., Cavanagh et al., 2011; Forstmann, Anwander, et al., 2010).

Furthermore, multiple routes to decision-threshold modulation have been identified, thereby demonstrating the value of this modeling approach for managing problems of the context sensitivity of cognitive function, causal ambiguity, and the task-impurity problem. On one hand, decision threshold in the speed-accuracy trade-off is modulated by changes in the functional connectivity between presupplementary motor area and striatum (Forstmann, Anwander, et al., 2010). On the other hand, neural-network modeling validated by studies of PD patients implanted with a deep-brain stimulator (Frank, Samanta, Moustafa, & Sherman, 2007) suggests that the STN is implicated in raising the decision threshold if there is conflict between two options associated with similar rewards. This result was further corroborated by Cavanagh et al. (2011), who found that trial-to-trial variations in frontal theta power (as measured by electroencephalography, EEG, as a measure of response conflict; Cavanagh, Zambrano-Vazquez, & Allen, 2012) is correlated with an increase in decision threshold during high-conflict trials. As predicted, this relationship was reversed when STN function was disrupted by deep-brain stimulation in PD patients. When deep-brain stimulators were turned off, patients exhibited the same conflict-induced regulation of decision threshold as a function of cortical theta. Similarly, intraoperative recordings of STN field potentials and neuronal spiking showed that STN activity responds to conflict during decision making, and is predictive of more accurate but slower decisions, as expected as a result of threshold regulation (Cavanagh et al., 2011; Zaghloul et al., 2012; Zavala et al., 2013). These results provide a computational cognitive explanation for the clinical symptom of impulsivity observed in PD patients receiving deep-brain stimulation (Bronstein et al., 2011; Frank, Samanta, et al., 2007; Hälbig et al., 2009).

Application to computational psychiatry. Despite its long history, the DDM has been applied to the study of psychopathology only recently. For example, threat/no-threat categorization tasks (e.g., "Is this word threatening or not?") are used in anxiety research to explore biases to threat responses. Subjects with high anxiety are more likely to classify a word as threatening than are low-anxiety subjects, although the explanation of this bias is unclear. One hypothesis assumes that this behavior results from an increased response bias toward threatening words in anxious people (Becker & Rinck, 2004; Manguno-Mire, Constans, & Geer, 2005; Windmann & Krüger, 1998). Using DDM analysis, White (2009) showed that instead of a response bias (or a shifted starting point in DDM terminology), anxious people showed a perceptual bias toward classifying threatening words, as indicated by an increased DDM drift rate.

In a recent review article, White, Ratcliff, Vasey, and McKoon (2010) used this case study to highlight the potential of the DDM to elucidate research into mental illness. Note that in this study, the authors did not hypothesize about the underlying neural cause of this threat bias. Although there is some evidence that bias in decision making is correlated with activity in the parietal network (Forstmann, Brown, Dutilh, Neumann, & Wagenmakers, 2010), this was not tested in respect to threatening words. Ultimately, we suggest that this research strategy should be applied to infer neural correlates of psychological DDM decision-making parameters using functional methods such as fMRI and employing modeling techniques at multiple levels of analysis.

The DDM has also been successfully used to show that ADHD subjects were less able to raise their decision threshold when accuracy demands were high (Mulder et al., 2010). It is interesting that the amount by which ADHD subjects did not modulate their decision threshold correlated strongly with patients' impulsivity/hyperactivity rating. Moreover, this correlation was specific to impulsivity and not inattentiveness. Note that in this case, the use of the *DSM* category (ADHD) may have obscured a more robust transdiagnostic association between decision-threshold modulation and hyperactivity, and "hyperactivity" itself may mask a variety of different causal processes.

A recent study by Pe, Vandekerckhove, and Kuppens (2013) showed that the DDM could also be used to explain previously conflicting reports on the influence of negative distractors on the emotional flanker task in depressed patients. Specifically, depression and rumination (a core symptom of depression) were associated with enhanced processing of negative information. These results further support the theory that depression is characterized by biased processing of negatively connotated information. Critically, this result could not be established by analyzing mean RT or accuracy alone, thereby demonstrating the enhanced sensitivity to cognitive behavior of computational models.

In sum, SSMs show great promise as a tool for computational psychiatry. In helping to map out the complex interplay of cognitive processes and their neural correlates in mental illness, such models can play a role in resolving task impurity and other forms of causal ambiguity, identifying and measuring cognitive impairments, and associating such impairments with both symptoms and neural correlates. However, their applicability depends on the ability to accurately estimate them to construct individual computational MPs (CMPs). Such CMPs are parameter profiles that represent an individual's functioning as measured by the specific parameters that make up the profile and derived from fitting the model to task-performance data. In the next section, we review different (Level 3) parameter-estimation techniques with a special focus on Bayesian methods that are usable for estimating parameters in the DDM and for generating individual CMPs. Finally, once SSMs can be fit accurately, we move on to identify (Level 4) clustering methods that can be used in a Bayesian framework to identify meaningful clusters of individuals, given their cognitive profiles (CMPs).

Level 3: Parameter estimation

It is critical to have robust and sensitive estimation methods to identify computational parameters in a variable clinical population with the DDM. In the following, we describe traditional parameter-estimation methods and their pitfalls. We then explain how Bayesian estimation provides a complete framework that avoids these pitfalls.

Random versus fixed parameters across groups of subjects. Fitting of computational models traditionally is treated as an optimization problem in which an objective function is minimized. Psychological experiments often test multiple subjects on the same behavioral task. Models are then fit either to individual subjects or to the aggregated group data. Both approaches are not ideal. When models are fit to individual subjects, we neglect any similarity the parameters are likely to have. Although we do not necessarily have to make use of this property to make useful inferences if we have lots of data, the ability to infer subject parameters on the basis of the estimation of other subjects generally leads to more accurate parameter recovery (Wiecki, Sofer, & Frank, 2013) in cases in which little data are available, as is often the case in clinical and neurocognitive experiments. One alternative is to aggregate all subject data into one meta-subject and estimate one set of parameters for the whole group. Although useful in some settings, this approach is unsuited for the setting of computational psychiatry, given that individual differences play a huge role.

Hierarchical Bayesian models. Statistics and machine learning have developed efficient and versatile Bayesian methods to solve various inference problems (Poirier, 2006). They more recently have seen wider adoption in applied fields such as genetics (Stephens & Balding, 2009) and psychology (e.g., Clemens, De Vrijer, Selen, Van Gisbergen, & Medendorp, 2011). One reason for this Bayesian revolution is the ability to quantify the certainty one has in a particular estimation. Moreover, hierarchical Bayesian models provide an elegant solution to the problem of estimating parameters of individual subjects outlined earlier (viz., the problem of neglecting similarities of parameters across subjects). Under the assumption that subjects within each group are similar to each other, but not identical, a hierarchical model can be constructed in which individual parameter estimates are constrained by group-level distributions (Nilsson, Rieskamp, & Wagenmakers, 2011; Shiffrin, Lee, & Kim, 2008), and more so if group members are similar to each other.

Thus, hierarchical Bayesian estimation leverages similarity between individual subjects to share statistical power and increase sensitivity in parameter estimation. However, note that in our computational-psychiatry application, the homogeneity assumption that all subjects come from the same normal distribution is almost certainly violated (see earlier discussion). For example, differences between subgroups of ADHD subjects would be decreased as the normality assumption pulls them closer together. To deal with the heterogeneous data often encountered in psychiatry, we discuss mixture models in a later section. A detailed description of the mathematical details and inference methods of Bayesian statistics relevant for this endeavor can be found in the Bayesian Inference section in the Supplemental Material.

Level 4: Supervised and unsupervised learning

Given that parameters have been estimated, or even given behavioral statistics alone, how can we group individuals into clusters that might be relevant for diagnostic categories or treatments? Bayesian clustering algorithms are particularly relevant to our objective, given that they (a) deal with the heterogeneity encountered in computational psychiatry and (b) have the potential to bootstrap new classifications on the basis of measurable, quantitative, computational endophenotypes. Because we are describing a toolbox using hierarchical Bayesian estimation techniques, we focus this section on mixture models, given that they are easily integrated into this framework. Where possible, we highlight connections to more traditional clustering methods (e.g., "*k*-means").

Gaussian mixture models. Gaussian mixture models (GMMs) assume parameters to be distributed according to one of several Gaussian distributions (i.e., clusters). Specifically, given the number of clusters *k*, each cluster mean and variance is estimated from the data. This type of model is capable of solving our earlier identified problem of assuming heterogeneous subjects to be normally distributed: By allowing individual subject parameters to be assigned to different clusters, we allow estimation of different subgroups in our patient and HC populations. Note, however, that the number *k* of how many clusters should be estimated must be specified a priori in a GMM and remain fixed for the course of the estimation. This is problematic, given that we do not necessarily know how many subgroups to expect in advance. Bayesian nonparametrics solve this issue by inferring the number of clusters from data. Dirichlet processes GMMs (DPGMMs) belong to the class of Bayesian nonparametrics (Antoniak, 1974). They can be viewed as a variant of GMMs with the critical difference that they infer the number of clusters from the data (for a review, see Gershman & Blei, 2012). An arguably simpler alternative, however, is to run multiple clusterings tested with different numbers of clusters and perform model comparison, as we discuss next.

Model comparison. Model comparison provides measures to evaluate how well a model can explain the data while at the same time penalizing model complexity. Measures such as the Bayesian information criterion (mathematical details can be found in the Model Comparison section in the Supplemental Material) can be used to choose the GMM with the least number of clusters that still provide a good fit to the data. Moreover, model comparison is used to select between computational cognitive models that often allow formulation of several plausible accounts of cognitive behavior. Of particular note are Bayes factors that measure the evidence of a particular model in comparison with other, competing models (Kass & Raftery, 1993). More recent, and highly relevant to the field of computational psychiatry, these methods have been extended to provide proper random-effects inference on model structure in heterogeneous populations (Stephan, Penny, Daunizeau, Moran, & Friston, 2009).

Example Applications

In this last section, we provide a proof of concept by demonstrating how the earlier described techniques (Levels 1–4) can be combined to (a) recover clusters associated with age, on the basis of CMPs as extracted by the DDM; and (b) predict brain state (deep-brain stimulation on/off).

Supervised and unsupervised learning of age

To demonstrate the concepts presented here, we reanalyzed a data set collected and published by Ratcliff et al. (2010). The data set consists of two groups of human subjects, young (mean age 20.8) and old (mean age 68.6), tested on three different tasks: (a) a numerosity-discrimination task that involved estimation of whether the number of asterisks presented on the screen was more or less than 50 (such that trials with close to 50 asterisks were harder than were those with far fewer or far greater), (b) a lexical decision task that required subjects to decide whether a presented string of letters is an existing word of the English language, and (c) a memory-recognition task that presented words to be remembered in a training phase that were subsequently tested for recall together with distractor words. Details of the tasks (including the conditions tested), subject characteristics, and DDM analyses can be found in the original publication (Ratcliff et al., 2010).

We used the hierarchical DDM (HDDM) toolbox (Wiecki et al., 2013) to perform hierarchical Bayesian estimation of DDM parameters from subjects' RT and choice data without taking the different groups into account. We concatenated the DDM parameters of each subject in three tasks into one 22-dimensional CMP.

We next performed factor analysis on the CMP vectors. Factor analysis is a statistical technique that uses correlations between parameters to find latent variables (called factors). Intuitively, highly correlated parameters will be loaded onto the same factor. As shown in the factor-loading matrix in Figure 3, DDM parameters related to processing capability (i.e., drift rate) in the three tasks are loaded onto the first four factors, whereas nondecision times and thresholds in the three tasks are loaded onto Factors 5 and 6, respectively. Thus, instead of the 22 original dimensions, we are able to describe the cognitive variables of individuals using six latent factors.

Classification of impairments and dysfunctions based on CMPs is a critical requirement for the clinical application of computational psychiatry. Although classification of age might not have clinical relevancy, it provides an ideal testing environment because age is objectively measurable (as opposed to, e.g., SZ, as described earlier). To

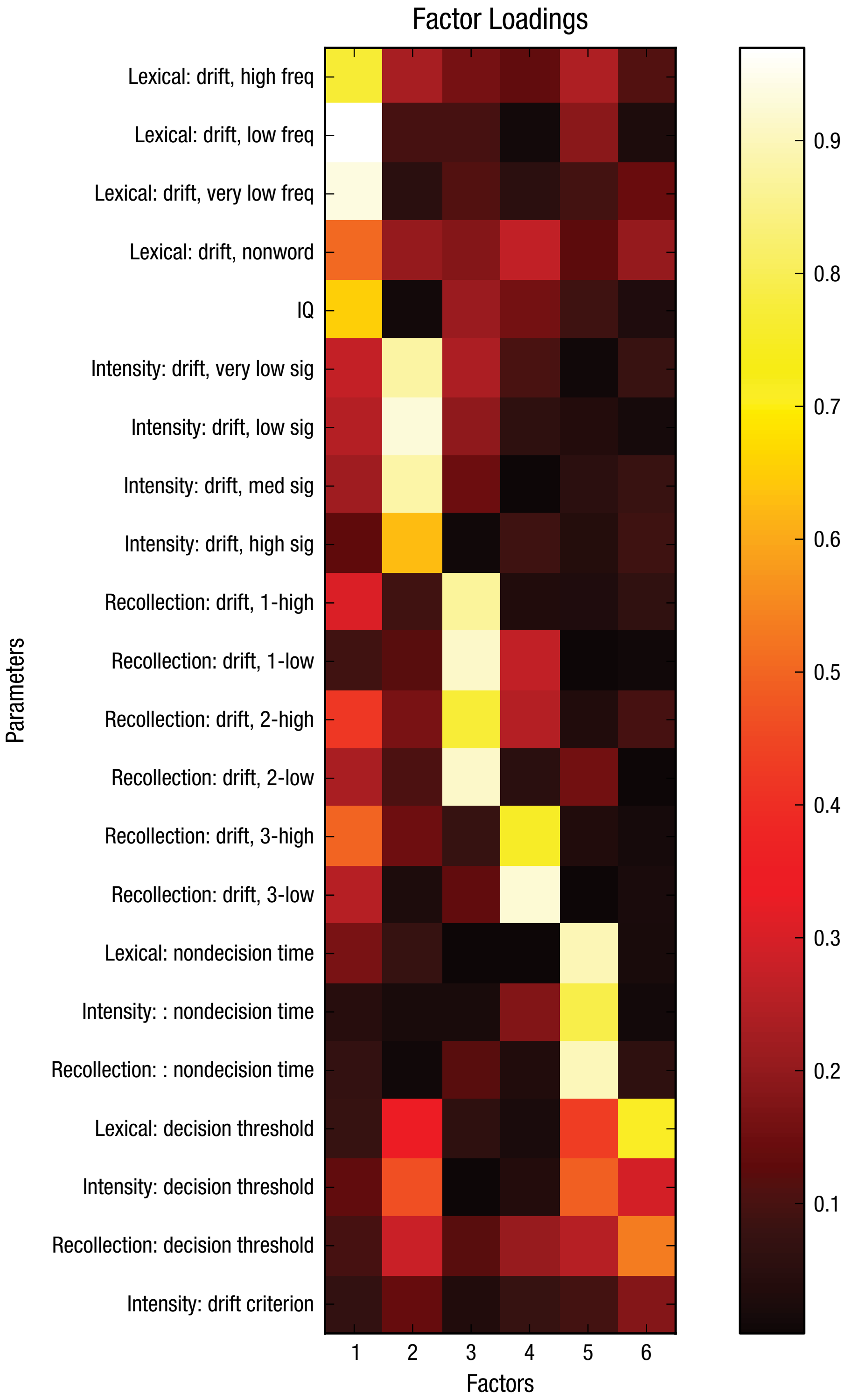

Fig. 3. Factor-loading matrix. Drift-diffusion model parameters of three tasks are presented along the *y*-axis; the extracted factors are distributed along the *x*-axis. Color codes indicate loading strengths. See the textual discussion for more details. freq = frequency; sig = significance.

classify young versus old, we employed logistic regression (using Level-2 regularization) on a subset of the data and evaluated its prediction accuracy using held-out data (by using cross-validation). Classification performance was very high (up to 95% accuracy; not shown), which demonstrated that cognitive tasks show great potential

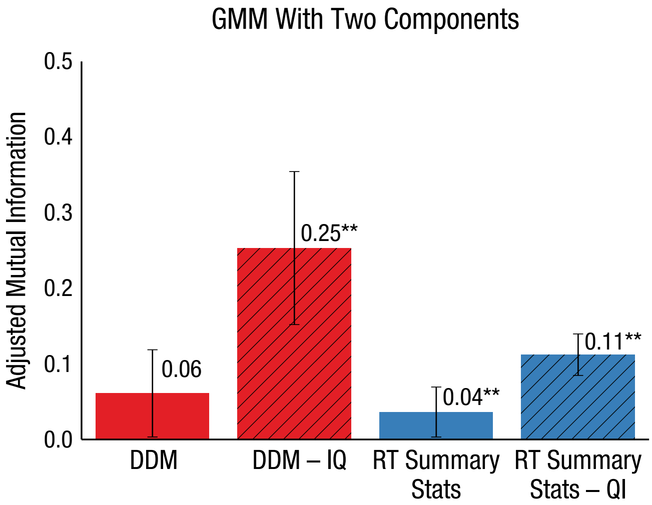

Fig. 4. Results: adjusted mutual information scores (higher is better, where 1 would mean perfect label recovery and 0 would mean chance level) for age after estimation of a Gaussian mixture model (GMM) with two components on drift-diffusion-model (DDM) factors (see text for more details on the factor analysis) and on DDM factors after the contribution of IQ was regressed out. Error bars represent standard deviations assessed via bootstrap. Asterisks denote significantly higher chance performance (**p < .01). RT = response time.

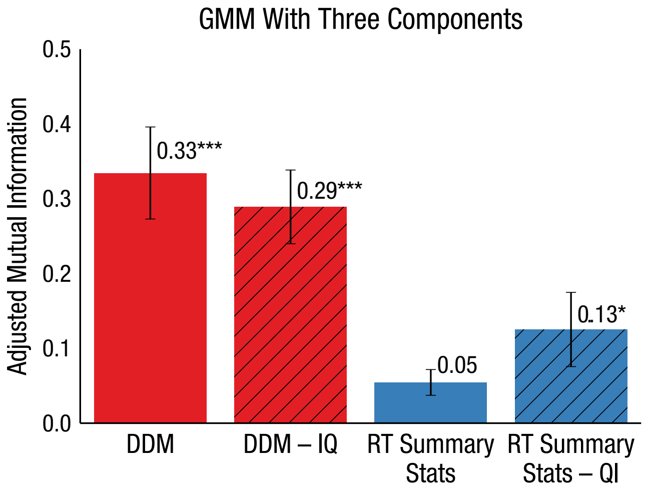

Fig. 5. Results: adjusted mutual information scores (higher is better, where 1 would mean perfect label recovery and 0 would mean chance level) for age after estimation of a Gaussian mixture model (GMM) with three components on drift-diffusion-model (DDM) factors (see text for more details on the factor analysis) and on DDM factors after the contribution of IQ was regressed out. Error bars represent standard deviations assessed via bootstrap. Asterisks denote significantly higher chance performance (*p < .05, ***p < .001). RT = response time.

for classifying differences in brain functioning. In this case, there was no benefit to using DDM parameters compared with using summary statistics on RT and accuracy, given that the differences in behavioral profiles between subjects with large differences in age were quite stark. There are several examples in which usage of a computational model does yield a significant increase in classification accuracy (see later discussion; also see Brodersen et al., 2013) and may be more likely to do so if the patterns are more nuanced.

When these techniques are used to classify a mental illness such as SZ, there is concern about the validity of our labels. If SZ does not represent a homogeneous, clearly defined group of individuals but, rather, patients with various cognitive and mental abnormalities, how could we expect a classifier to predict such an elusive, ill-defined label? One potential way to deal with this problem is to use an unsupervised clustering algorithm to find a new grouping that is hopefully more sensitive to the neurocognitive deficits (Fair et al., 2012). As a proof of principle, we tested how well GMM clustering could recover age-groupings in an unsupervised manner. Note that in a clinically more relevant setting, we would not necessarily know the correct grouping ahead of time. Figure 4 shows the adjusted mutual information (which is 1 if we perfectly recover the original grouping and 0 if we group by chance) for age when estimating two clusters based on six latent factors extracted using factor analysis (we did not include IQ in the factor analysis here). It is notable that the age cluster is not recovered at all when the DDM factors are used. Follow-up analysis suggests that the clustering selected by GMM picks up on some of the structure introduced by IQ (adjusted mutual information = 0.1; not shown). This indeed represents a potential problem for this unsupervised approach, given that there are many sources of individual variation, such as age, IQ, or education, we might not be interested in when wanting clusters sensitive to pathological sources of variation. To address this problem, we regressed the contribution of IQ out of every factor to remove this source of variation. Running GMM on these new regressed factors, we observed that the algorithm now clusters into different age-groups (adjusted mutual information is 0.25, which corresponds to an accuracy of approximately 75%). This might thus provide a viable technique in removing unwanted sources of interindividual variation, given that variables such as age, IQ, or education could just be regressed out before doing the clustering—if these nuisance variables are known and measured.

The main issue here is that multiple factors can contribute to clusterings of neurocognitive parameters. A different solution to this problem is presented in Figure 5, in which we estimated a GMM allowing for an additional cluster (three clusters total). As the figure shows, even when not regressing IQ out of the parameters, the clustering solution shows a clear sensitivity to age, albeit none to IQ. Moreover, the use of summary statistics on RT and accuracy (mean and standard deviation) alone did not achieve a comparable level of recovery with the GMM (see Figs. 3 and 4). We also performed model comparison using BIC (not shown) to find the best number of clusters when we successively tested different numbers of clusters. We found that adding more clusters monotonically decreased BIC thus favoring models with many clusters, despite the added complexity of these models. This might not be surprising given that there are many other individual differences beyond age and IQ that could affect group membership. It does represent a problem for this approach, however, given that it is not immediately clear what level of representation should be chosen if a purely unsupervised measure such as BIC does not provide guidance.

In conclusion, we demonstrated how computational modeling and latent variable models can be used to construct CMPs of individuals tested on multiple cognitive decision-making tasks. Using supervised machine-learning methods, we were able to achieve up to 95% accuracy in classifying young versus old age. Finally, after we regressed IQ out as a nuisance variable, unsupervised clustering was able to group young and old individuals on the basis of the structure of the CMP space.

Simulation experiment

Although the preceding example demonstrated a clear benefit in using the DDM for unsupervised clustering, the model parameters were less beneficial compared with simple behavioral summary statistics (RT and accuracy) when we performed supervised classification. This finding raises the question whether DDM parameters derived on the basis of behavioral measures alone can, in principle, provide a benefit in supervised learning over summary statistics. We thus performed a simple experiment in which we simulated data from the DDM generating two groups with 40 subjects each. The mean parameters of the two groups differed in threshold, drift rate, and nondecision time (exact values can be found in the Parameters Used in Simulation Study section in the Supplemental Material). We then recovered DDM parameters by estimating the HDDM (without allowing group to influence fit, which would be an unfair bias). Summary statistics consisted of mean and standard deviation of RT and accuracy. Figure 6 shows the area under the curve using logistic regression with Level-2 regularization in a 10-fold cross-validation. As the figure shows, for this parameter setting, the DDM-recovered parameters provide a large benefit over summary statistics. During the exploration of various generative parameter settings, however, we also found that other settings do not lead to an improvement, similar to the result obtained on the aging data set. Further research is necessary to establish conditions under which DDM provides a clear benefit over using the simpler summary statistics.

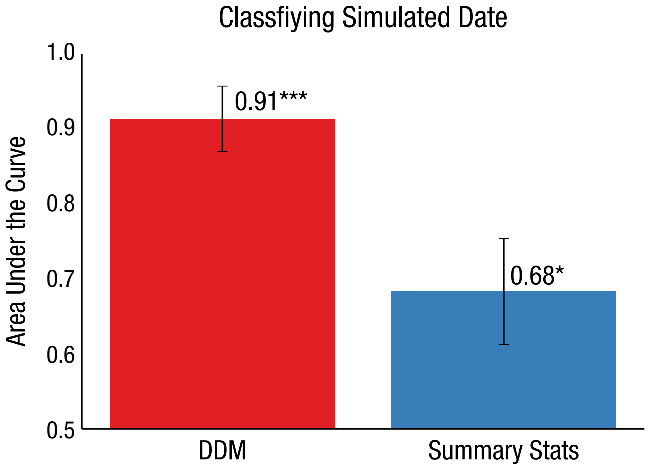

Fig. 6. Area under the receiver operating characteristic curve that relates to classification accuracy of simulated response-time data from the drift-diffusion model (DDM). DDM represents parameters recovered in a hierarchical DDM fit ignoring the group labels. Summary statistics are mean and standard deviation of response time and accuracy. Error bars represent standard deviations. Asterisks indicate accuracy significantly higher than chance (**p* < .05, ****p* < .001).

Predicting brain state on the basis of EEG

The previously discussed age example clearly demonstrated the potential of this approach in a data-driven, hypothesis-free manner. To complement this example, we tested whether it was possible, using computational methods, to classify patients' brain state using computational parameters related to measures of impulsivity. We reanalyzed a data set from our lab in which PD patients implanted with deep-brain stimulators in the STN were tested on a reward-based decision-making task (Cavanagh et al., 2011). STN deep-brain stimulation is very effective in treating the motor symptoms of the disease but can sometimes cause cognitive deficits and impulsivity (Bronstein et al., 2011; Hälbig et al., 2009). Prior work has shown that when faced with conflict between different reward values during decision making, HC subjects and patients off deep-brain stimulation adaptively slow down to make a more considered choice, whereas STN deep-brain stimulation induces fast impulsive actions. In this study, we showed that the degree of RT slowing for high-conflict trials was related to the degree to which frontal theta power increased. DDM model fits revealed that theta power increases were specifically related to an increase in decision threshold, thereby leading to more cautious but accurate responding, whereas deep-brain stimulation

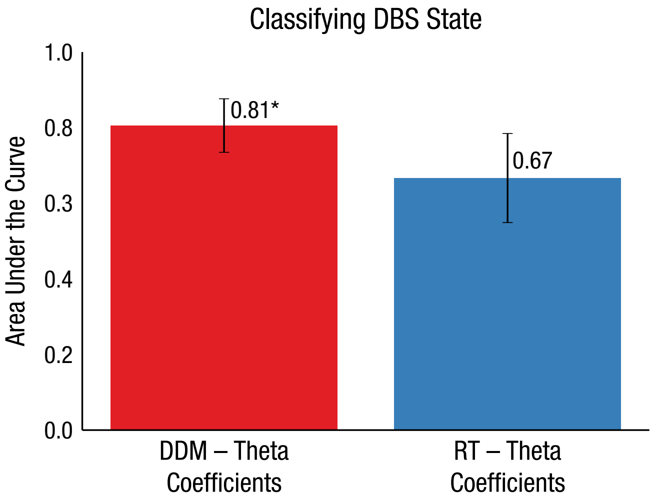

Fig. 7. Out-of-sample classification accuracy using logistic regression to deep-brain-stimulation (DBS) state comparing drift-diffusion-model (DDM) coefficients and using regression between response time (RT) and theta power. Error bars indicate standard deviations based on a bootstrap. The asterisk encodes significance (*$p < .05$).

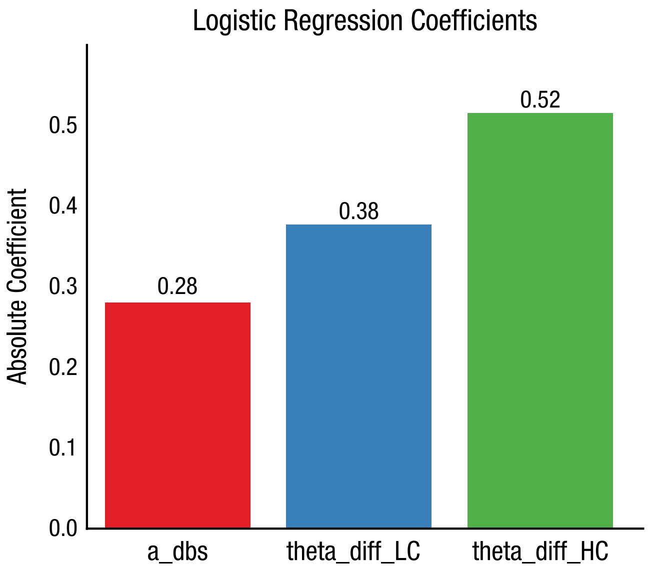

Fig. 8. Results: absolute coefficients of logistic regression model using three predictors. Intuitively, the higher the coefficient, the more it contributes to separability of deep-brain-stimulation (DBS) state. The difference in threshold between DBS on and off is a_dbs; theta_diff_LC and theta_diff_HC are the differences in trial-by-trial regression coefficients between theta power (as measured via electroencephalography) and decision threshold for low- and high-conflict trials, respectively.

prevented patients from increasing their threshold despite increases in cortical theta, which led to impulsive choice.

These findings lend support to a computational hypothesis based on a variety of data across species regarding the neural mechanisms for decision-threshold regulation. However, the findings were significant at the group level. Here, we tested whether we could classify individual patients' deep-brain-stimulation status knowing only their DDM parameters (estimated from RT and choice data). We also included as a predictor the degree to which frontal theta modulated decision threshold (effectively, another DDM parameter). Specifically, we used logistic regression with Level-2 regularization and cross-validation. The features for the classifier were the difference in thresholds in the two brain states (on and off deep-brain stimulation) and the difference in the theta-threshold regression coefficients in high- and low-conflict trials (on and off deep-brain stimulation). The classifier tries to predict which brain state a new subject is in on the basis of these difference parameters without informing it as to which one corresponds to the on or off state. We randomly sampled binary labels for each subject. The label indicated whether the features were coded relative to the on or off state. Intuitively, if the label were 0 for a subject, the features would contain the change in regression coefficients (theta_diff_LC for low conflict and theta_diff_HC for high conflict) and threshold (a_dbs) when going from deep-brain stimulation on to off. Conversely, if the label were 1, the features would contain the change in regression coefficients and threshold when going from deep-brain stimulation off to on. The job of the classifier then becomes the classification of whether an individual is in the deep-brain stimulation on or off state on the basis of the change in coefficients. The features based on raw RT data were created in a similar manner: Instead of using the regression coefficients of the influence of theta on decision threshold, we included the influence of theta directly on RT in low and high conflict (shown to be significantly correlated in Cavanagh et al., 2011) as well as the difference in mean RT between deep-brain stimulation on and off.

As shown in Figure 7, use of the DDM analysis greatly improved classification accuracy. It is interesting that of all the parameters fed into the classifier, the degree to which theta related to threshold adjustments in high-conflict trials was most predictive of deep-brain stimulation state (see Fig. 8 for absolute coefficients of the logistic regression model using three predictors). This result is consistent with that obtained in Cavanagh et al. (2011) but extends it to show how an individual patient's brain state, as a biomarker of impulsivity, can be diagnosed.

We thus demonstrated that this DDM analysis can be combined with brain measures (here EEG, but other measures, such as fMRI, are just as viable) to predict very specific changes in brain state. Critically, the influence of EEG on RT alone, although significant in Cavanagh et al. (2011), did not allow for the same accuracy as the DDM analysis. Moreover, this example shows the value of being hypothesis driven, given that this link between decision threshold and theta in high-conflict trials (which was recovered as the most discriminative feature) was suggested by earlier, biologically plausible modeling efforts (Wiecki & Frank, 2013).

Although aggregate performance scores (i.e., MPs, such as mean accuracy or mean RT) could, in principle, be used for classification and clustering, there are some unique advantages of using CMPs:

- Computational models distill domain knowledge of the cognitive processes underlying task performance. For this reason, they can be seen as feature extraction methods that reduce nuisance variables and find a process-based representation of cognitive ability and, thus, make it easier for the classifier to separate different groups.
- Computational modeling can help with the task-impurity problem. Aggregate performance scores summarize the contribution of a mixture of cognitive processes involved in a task. Computational models try to deconstruct behavior into its individual components and identify separable cognitive processes.
- Neurocognitive models often assume cognitive processes to be implemented by certain networks of the brain. For this reason, a computational parameter identified to have predictive power can be linked much easier to neural processes than aggregate performance scores.

Applications and Challenges

How could this research program improve mental-health diagnosis, treatment, and research? The ultimate hope is that psychiatric diagnosis could move away from a symptom-based classification of mental illness and instead use quantifiable biomarkers. CMPs could contribute to this by quantifying subjects' cognitive abilities in terms of psychological-process variables that describe the efficacy of their neural circuitry.

Psychiatric drugs, as well as other forms of treatment, including deep-brain stimulation, have a high degree of variability in their efficacy across individuals. By identifying pathological cascades and how they interact with treatment, we might be able to predict which form of treatment would be effective for an individual and optimize treatment variables.

With regard to clinical research, computational psychiatry can provide tools to link clinical symptoms to neurocognitive dysfunction that can open the door to a deeper level of understanding as well as provide novel targets for future studies into the causes of mental illness. For pharmacological research, assessment of a drug mechanism and its efficacy by clinical ratings alone is often noisy, hard to interpret, and biased as a result of the placebo effect. More objective and quantitative measures of neurocognitive function are likely to improve on these current issues. Moreover, many psychiatric drugs fail in Phase 3 clinical trials even though they show promising results for a small subset of enrolled patients. If that subset could be identified by cognitive testing, the output of the drug-discovery pipeline could be enhanced.

Although the potential fruits of this research program are thus promising, the expected challenges to be overcome are nevertheless substantial. We cannot rely on *DSM* categories or a foundational understanding of the brain to bootstrap a new system in which to redefine mental illness. Among the main challenges is finding a good description of normal and abnormal cognitive function. Are there distinct clusters of cognitive dysfunction (and if so, how many), or is there a continuum with an arbitrary threshold on where mental illness begins? This article provides an example for how regressing out IQ can allow for better classification of age. In more complex psychiatric conditions, we clearly may not always have access to variables that affect clustering of behavioral phenotypes in ways over which we would like to abstract.

Although the new transdimensional approach of RDoC by the National Institute of Mental Health is very promising, it must be open to additional levels of descriptions, such as the neurocognitive computations of the brain. Computational psychiatry could then be embedded in this framework and translate neurocognitive research findings to other domains, including genetics, neuroscience, and clinical psychology.

Conclusions

In the light of the crisis in mental-health research and practice and the widely recognized problems with conventional psychiatric classification based on the *DSM*, computational psychiatry is an emerging field that shows great promise for pursuing research aimed at understanding mental illness. Computational psychiatry provides powerful conceptual and methodological resources that enable management of the various features of mental illness and the various challenges with which researchers must cope. More specific, by fitting computational models to behavioral data, we can estimate computational parameters and construct CMPs that provide measures of functioning in one or another cognitive domain. Such measures are potentially of value in research contexts previously organized around symptom-based classification as implemented by the *DSM*. CMPs may function as both more precise targets of research and more powerful explanatory resources for understanding individual differences, significant groupings, dynamic interactivity, and hierarchical organization of the brain.

Decision making appears to provide a good framework for studying mental illness, given that many disorders show abnormalities in core decision-making processes. Strategically designed task batteries can provide the behavioral basis for studying such abnormalities. SSMs have a good track record in describing individual differences in decision making and can be linked to neuronal processes. Hierarchical Bayesian estimation provides a compelling toolbox to fit these models directly to data because it (a) provides an uncertainty measure, (b) allows estimation of individual- and group-level parameters simultaneously, (c) allows for direct model comparison, and (d) enables deconstruction of symptoms by identifying latent clusters that correspond to different causal mechanisms. For example, impulsivity is a core symptom of many mental disorders, such as ADHD, OCD, Tourette's syndrome, and substance-abuse and eating disorders (Robbins et al., 2012). Computational cognitive models have already started to deconstruct this broadly defined behavioral symptom and have identified separate pathways that can all lead to alterations in impulse control (Dalley, Everitt, & Robbins, 2011), including reduced motor inhibition (Chamberlain et al., 2006; Chamberlain et al., 2008), early temporal discounting of future rewards, insensitivity toward negative relative to positive outcomes (Cockburn & Holroyd, 2010; Frank, Santamaria, O'Reilly, & Willcutt, 2007), or an inability to adjust the decision threshold appropriately (Cavanagh et al., 2011; Frank, Samanta, et al., 2007; Mulder et al., 2010).

Ultimately, the hope is to find novel ways to describe and assess mental illness on the basis of objective computational neurocognitive parameters rather than the current subjective symptom-based approach. The bottom line is that computational psychiatry provides a combination of computational tools and strategies that are potentially powerful enough to underwrite a research program that will lead to a new level of understanding of mental illness and to new ways to describe, investigate, and assess mental illness on the basis of identifiable and reproducible neurocognitive CMPs.

Author Contributions

All research was performed by T. V. Wiecki under supervision of M. J. Frank. The paper was written by T. V. Wiecki, M. J. Frank, and J. Poland.

Acknowledgments

The authors are grateful to Roger Ratcliff for generously providing the aging data set and for useful discussions.

Declaration of Conflicting Interests

M. J. Frank consults for F. Hoffman–La Roche Pharmaceuticals using computational-psychiatry methods (but not those reported in this article).

Funding

This research was supported by National Institute of Mental Health Grant R01 MH080066-01.

Supplemental Material

Additional supporting information may be found at http://cpx.sagepub.com/content/by/supplemental-data

References

Aichert, D. S., Wöstmann, N. M., Costa, A., Macare, C., Wenig, J. R., Möller, H.-J., . . . Ettinger, U. (2012). Associations between trait impulsivity and prepotent response inhibition. *Journal of Clinical and Experimental Neuropsychology*, *34*, 1016–1032.

American Psychiatric Association. (2013). *Diagnostic and statistical manual of mental disorders* (5th ed.). Arlington, VA: American Psychiatric Publishing.

Antoniak, C. E. (1974). Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems. *The Annals of Statistics*, *2*, 1152–1174.

Badcock, J. C., Michie, P. T., Johnson, L., & Combrinck, J. (2002). Acts of control in schizophrenia: Dissociating the components of inhibition. *Psychological Medicine*, *32*, 287–297.

Becker, E., & Rinck, M. (2004). Sensitivity and response bias in fear of spiders. *Cognition & Emotion*, *18*, 961–976.

Bellgrove, M. A., Chambers, C. D., Vance, A., Hall, N., Karamitsios, M., & Bradshaw, J. L. (2006). Lateralized deficit of response inhibition in early-onset schizophrenia. *Psychological Medicine*, *36*, 495–505.

Bitzer, S., Park, H., Blankenburg, F., & Kiebel, S. J. (2014). Perceptual decision making: Drift-diffusion model is equivalent to a Bayesian model. *Frontiers in Human Neuroscience*, *8*, 102. Retrieved from http://journal.frontiersin.org/Journal/10.3389/fnhum.2014.00102/full

Brodersen, K. H., Deserno, L., Schlagenhauf, F., Lin, Z., Penny, W. D., Buhmann, J. M., & Stephan, K. E. (2013). Dissecting psychiatric spectrum disorders by generative embedding. *NeuroImage: Clinical*, *4*, 98–111.

Bronstein, J. M., Tagliati, M., Alterman, R. L., Lozano, A. M., Volkmann, J., Stefani, A., . . . DeLong, M. R. (2011). Deep brain stimulation for Parkinson disease: An expert consensus and review of key issues. *Archives of Neurology*, *68*, 165.

Buckholtz, J. W., & Meyer-Lindenberg, A. (2012). Psychopathology and the human connectome: Toward a transdiagnostic model of risk for mental illness. *Neuron*, *74*, 990–1004. doi:10.1016/j.neuron.2012.06.002

Cavanagh, J. F., Wiecki, T. V., Cohen, M. X., Figueroa, C. M., Samanta, J., Sherman, S. J., & Frank, M. J. (2011). Subthalamic nucleus stimulation reverses mediofrontal influence over decision threshold. *Nature Neuroscience*, *14*, 1462–1467.

Cavanagh, J. F., Zambrano-Vazquez, L., & Allen, J. J. B. (2012). Theta lingua franca: A common mid-frontal substrate for action monitoring processes. *Psychophysiology*, *49*, 220–238.

Chamberlain, S. R., Fineberg, N. A., Blackwell, A. D., Robbins, T. W., & Sahakian, B. J. (2006). Motor inhibition and cognitive flexibility in obsessive-compulsive disorder and trichotillomania. *American Journal of Psychiatry*, *163*, 1282–1284.

Chamberlain, S. R., Menzies, L., Hampshire, A., Suckling, J., Fineberg, N. A., del Campo, N., . . . Sahakian, B. J. (2008). Orbitofrontal dysfunction in patients with obsessive-compulsive disorder and their unaffected relatives. *Science*, *321*, 421–422.

Clemens, I. A. H., De Vrijer, M., Selen, L. P. J., Van Gisbergen, J. A. M., & Medendorp, W. P. (2011). Multisensory processing in spatial orientation: An inverse probabilistic approach. *Journal of Neuroscience*, *31*, 5365–5377.

Cockburn, J., & Holroyd, C. B. (2010). Focus on the positive: Computational simulations implicate asymmetrical reward prediction error signals in childhood attention-deficit/hyperactivity disorder. *Brain Research*, *1365*, 18–34.

Collins, A. G. E., & Frank, M. J. (2012). How much of reinforcement learning is working memory, not reinforcement learning? A behavioral, computational, and neurogenetic analysis. *European Journal of Neuroscience*, *35*, 1024–1035. doi:10.1111/j.1460-9568.2011.07980.x

Cressey, D. (2011, June). Psychopharmacology in crisis. *Nature*. Retrieved from http://www.nature.com/news/2011/110614/full/news.2011.367.html

Dalley, J. W., Everitt, B. J., & Robbins, T. W. (2011). Impulsivity, compulsivity, and top-down cognitive control. *Neuron*, *69*, 680–694.

Deneve, S. (2008). Bayesian spiking neurons: I. Inference. *Neural Computation*, *20*, 91–117.

Durston, S., Fossella, J. A., Mulder, M. J., Casey, B. J., Ziermans, T. B., Vessaz, M. N., & Van Engeland, H. (2008). Dopamine transporter genotype conveys familial risk of attention-deficit/hyperactivity disorder through striatal activation. *Journal of the American Academy of Child and Adolescent Psychiatry*, *47*, 61–67.

Elvevåg, B., & Goldberg, T. E. (2000). Cognitive impairment in schizophrenia is the core of the disorder. *Critical Reviews in Neurobiology*, *14*, 1–21.

Fair, D. A., Bathula, D., Nikolas, M. A., & Nigg, J. T. (2012). Distinct neuropsychological subgroups in typically developing youth inform heterogeneity in children with ADHD. *Proceedings of the National Academy of Sciences, USA*, *109*, 6769–6774.

Forstmann, B. U., Anwander, A., Schäfer, A., Neumann, J., Brown, S., Wagenmakers, E.-J., . . . Turner, R. (2010). Cortico-striatal connections predict control over speed and accuracy in perceptual decision making. *Proceedings of the National Academy of Sciences, USA*, *107*, 15916–15920.

Forstmann, B. U., Brown, S., Dutilh, G., Neumann, J., & Wagenmakers, E.-J. (2010). The neural substrate of prior information in perceptual decision making: A model-based analysis. *Frontiers in Human Neuroscience*, *4*, 40. Retrieved from http://journal.frontiersin.org/Journal/10.3389/fnhum.2010.00040/full

Frank, M. J., Doll, B. B., Oas-Terpstra, J., & Moreno, F. (2009). Prefrontal and striatal dopaminergic genes predict individual differences in exploration and exploitation. *Nature Neuroscience*, *12*, 1062–1068.

Frank, M. J., Samanta, J., Moustafa, A. A., & Sherman, S. J. (2007). Hold your horses: Impulsivity, deep brain stimulation, and medication in Parkinsonism. *Science*, *318*, 1309–1312.

Frank, M. J., Santamaria, A., O'Reilly, R. C., & Willcutt, E. (2007). Testing computational models of dopamine and noradrenaline dysfunction in attention deficit/hyperactivity disorder. *Neuropsychopharmacology*, *32*, 1583–1599.

Friston, K. J., Harrison, L., & Penny, W. (2003). Dynamic causal modelling. *NeuroImage*, *19*, 1273–1302.

Fukumoto-Motoshita, M., Matsuura, M., Ohkubo, T., Ohkubo, H., Kanaka, N., Matsushima, E., . . . Matsuda, T. (2009). Hyperfrontality in patients with schizophrenia during saccade and antisaccade tasks: A study with fMRI. *Psychiatry and Clinical Neurosciences*, *63*, 209–217.

Gershman, S. J., & Blei, D. M. (2012). A tutorial on Bayesian nonparametric models. *Journal of Mathematical Psychology*, *56*, 1–12.

Gold, J. I., & Shadlen, M. N. (2002). Banburismus and the brain: Decoding the relationship between sensory stimuli, decisions, and reward. *Neuron*, *36*, 299–308.

Gold, J. I., & Shadlen, M. N. (2007). The neural basis of decision making. *Annual Review of Neuroscience*, *30*, 535–574.

Gold, J. M., Waltz, J. A., Matveeva, T. M., Kasanova, Z., Strauss, G. P., Herbener, E. S., . . . Frank, M. J. (2012). Negative symptoms and the failure to represent the expected reward value of actions. *Archives of General Psychiatry*, *69*, 129–138.

Gold, J. M., Waltz, J. A., Prentice, K. J., Morris, S. E., & Heerey, E. A. (2008). Reward processing in schizophrenia: A deficit in the representation of value. *Schizophrenia Bulletin*, *34*, 835–847.

Hälbig, T. D., Tse, W., Frisina, P. G., Baker, B. R., Hollander, E., Shapiro, H., . . . Olanow, C. W. (2009). Subthalamic deep brain stimulation and impulse control in Parkinson's disease. *European Journal of Neurology*, *16*, 493–497.

Hallett, P. E. (1978). Primary and secondary saccades to goals defined by instructions. *Vision Research*, *18*, 1279–1296.

Huang, Y., & Rao, R. P. N. (2013). Reward optimization in the primate brain: A probabilistic model of decision making under uncertainty. *PLoS ONE*, *8*(1), e53344. Retrieved from http://www.plosone.org/article/info%3Adoi%2F10.1371%2Fjournal.pone.0053344

Huddy, V. C., Aron, A. R., Harrison, M., Barnes, T. R. E., Robbins, T. W., & Joyce, E. M. (2009). Impaired conscious and preserved unconscious inhibitory processing in recent onset schizophrenia. *Psychological Medicine*, *39*, 907–916.

Huys, Q. J. M., Eshel, N., O'Nions, E., Sheridan, L., Dayan, P., & Roiser, J. P. (2012). Bonsai trees in your head: How the Pavlovian system sculpts goal-directed choices by pruning decision trees. *PLoS Computational Biology*, *8*(3), e1002410. Retrieved from http://www.ploscompbiol.org/article/info%3Adoi%2F10.1371%2Fjournal.pcbi.1002410

Hyman, S. E. (2012). Revolution stalled. *Science Translational Medicine*, *4*(155), 1–5. doi:10.1126/scitranslmed.3003142

Insel, T., Cuthbert, B., Garvey, M., Heinssen, R., Pine, D. S., Quinn, K., . . . Wang, P. W. (2010). Research Domain Criteria (RDoC): Toward a new classification framework for research on mental disorders. *American Journal of Psychiatry*, *167*, 748–751.

Kass, R. E., & Raftery, A. E. (1993). *Bayes factors and model uncertainty* (Technical Report No. 571). Pittsburgh, PA: Carnegie Mellon University.

Maia, T. V. (2010). Two-factor theory, the actor-critic model, and conditioned avoidance. *Learning & Behavior*, *38*, 50–67.

Maia, T. V., & Frank, M. J. (2011). From reinforcement learning models to psychiatric and neurological disorders. *Nature Neuroscience*, *14*, 154–162.

Manguno-Mire, G. M., Constans, J. I., & Geer, J. H. (2005). Anxiety-related differences in affective categorizations of lexical stimuli. *Behaviour Research and Therapy*, *43*, 197–213.

Menzies, L., Achard, S., Chamberlain, S. R., Fineberg, N., Chen, C.-H., Del Campo, N., . . . Bullmore, E. (2007). Neurocognitive endophenotypes of obsessive-compulsive disorder. *Brain*, *130*, 3223–3236.

Montague, P. R., Dayan, P., & Sejnowski, T. J. (1996). A framework for mesencephalic dopamine systems based on predictive Hebbian learning. *Journal of Neuroscience*, *16*, 1936–1947.

Montague, P. R., Dolan, R. J., Friston, K. J., & Dayan, P. (2011). Computational psychiatry. *Trends in Cognitive Sciences*, *16*, 72–80.

Monterosso, J. R., Aron, A. R., Cordova, X., Xu, J., & London, E. D. (2005). Deficits in response inhibition associated with chronic methamphetamine abuse. *Drug and Alcohol Dependence*, *79*, 273–277.

Morein-Zamir, S., Fineberg, N. A., Robbins, T. W., & Sahakian, B. J. (2010). Inhibition of thoughts and actions in obsessive-compulsive disorder: Extending the endophenotype? *Psychological Medicine*, *40*, 263–272.

Mulder, M. J., Bos, D., Weusten, J. M. H., van Belle, J., van Dijk, S. C., Simen, P., . . . Durston, S. (2010). Basic impairments in regulating the speed-accuracy tradeoff predict symptoms of attention-deficit/hyperactivity disorder. *Biological Psychiatry*, *68*, 1114–1119.

Nigg, J. T. (2001). Is ADHD a disinhibitory disorder? *Psychological Bulletin*, *127*, 571–598.

Nigg, J. T., Wong, M. M., Martel, M. M., Jester, J. M., Puttler, L. I., Glass, J. M., . . . Zucker, R. A. (2006). Poor response inhibition as a predictor of problem drinking and illicit drug use in adolescents at risk for alcoholism and other substance use disorders. *Journal of the American Academy of Child and Adolescent Psychiatry*, *45*, 468–475.

Nilsson, H. K., Rieskamp, J., & Wagenmakers, E.-J. (2011). Hierarchical Bayesian parameter estimation for cumulative prospect theory. *Journal of Mathematical Psychology*, *55*, 84–93.

Nutt, D., & Goodwin, G. (2011). ECNP Summit on the future of CNS drug research in Europe 2011. *European Neuropsychopharmacology*, *21*, 495–499.

Oosterlaan, J., Logan, G., & Sergeant, J. (1998). Response inhibition in AD/HD, CD, comorbid AD/HD+CD, anxious, and control children: A meta-analysis of studies with the stop task. *Journal of Child Psychology and Psychiatry*, *39*, 411–425.

O'Reilly, R. C., Munakata, Y., Frank, M. J., & Hazy, T. E. Contributors. (2012). *Computational cognitive neuroscience* [Wiki book]. Retrieved from http://ccnbook.colorado.edu

Pe, M. L., Vandekerckhove, J., & Kuppens, P. (2013). A diffusion model account of the relationship between the emotional flanker task and rumination and depression. *Emotion*, *13*, 739–747.

Penadés, R., Catalán, R., Rubia, K., Andrés, S., Salamero, M., & Gastó, C. (2007). Impaired response inhibition in obsessive compulsive disorder. *European Psychiatry*, *22*, 404–410.

Poirier, D. J. (2006). The growth of Bayesian methods in statistics and economics since 1970. *Bayesian Analysis*, *1*, 969–979.

Poland, J., Von Eckardt, B., & Spaulding, W. (1994). Problems with the *DSM* approach to classifying psychopathology. In G. Graham & G. L. Stephens (Eds.), *Philosophical psychopathology* (pp. 235–260). Cambridge, MA: MIT Press.

Ratcliff, R., & McKoon, G. (2008). The diffusion decision model: Theory and data for two-choice decision tasks. *Neural Computation*, *20*, 873–922.

Ratcliff, R., & Rouder, J. N. (1998). Modeling response times for two-choice decisions. *Psychological Science*, *9*, 347–356.

Ratcliff, R., Thapar, A., & McKoon, G. (2010). Individual differences, aging, and IQ in two-choice tasks. *Cognitive Psychology*, *60*, 127–157.

Robbins, T. W., Gillan, C. M., Smith, D. G., de Wit, S., & Ersche, K. D. (2012). Neurocognitive endophenotypes of impulsivity and compulsivity: Towards dimensional psychiatry. *Trends in Cognitive Sciences*, *16*, 81–91.

Sahakian, B. J., Malloch, G., & Kennard, C. (2010). A UK strategy for mental health and wellbeing. *Lancet*, *375*, 1854–1855.

Schachar, R., & Logan, G. D. (1990). Impulsivity and inhibitory control in normal development and childhood psychopathology. *Developmental Psychology*, *26*, 710–720.

Schultz, W., Dayan, P., & Montague, P. R. (1997). A neural substrate of prediction and reward. *Science*, *275*, 1593–1599.

Shiffrin, R. M., Lee, M. D., & Kim, W. (2008). A survey of model evaluation approaches with a tutorial on hierarchical Bayesian methods. *Cognitive Science*, *32*, 1248–1284.

Smith, P. L., & Ratcliff, R. (2004). Psychology and neurobiology of simple decisions. *Trends in Neurosciences*, *27*, 161–168.

Sonuga-Barke, E. J. S. (2005). Causal models of attention-deficit/hyperactivity disorder: From common simple deficits to multiple developmental pathways. *Biological Psychiatry*, *57*, 1231–1238.

Stephan, K. E., Penny, W. D., Daunizeau, J., Moran, R. J., & Friston, K. J. (2009). Bayesian model selection for group studies. *NeuroImage*, *46*, 1004–1017.

Stephens, M., & Balding, D. J. (2009). Bayesian statistical methods for genetic association studies. *Nature Reviews Genetics*, *10*, 681–690.

Strauss, G. P., Frank, M. J., Waltz, J. A., Kasanova, Z., Herbener, E. S., & Gold, J. M. (2011). Deficits in positive reinforcement learning and uncertainty-driven exploration are associated with distinct aspects of negative symptoms in schizophrenia. *Biological Psychiatry*, *69*, 424–431.

Sutton, R. S., & Barto, A. G. (1998). *Reinforcement learning: An introduction*. Cambridge, MA: MIT Press.

Townsend, J. T., & Ashby, F. G. (1983). *The stochastic modeling of elementary psychological processes*. New York, NY: Cambridge University Press.

van Koningsbruggen, M. G., Pender, T., Machado, L., & Rafal, R. D. (2009). Impaired control of the oculomotor reflexes in Parkinson's disease. *Neuropsychologia*, *47*, 2909–2915.

Waltz, J. A., Frank, M. J., Robinson, B. M., & Gold, J. M. (2007). Selective reinforcement learning deficits in schizophrenia support predictions from computational models of striatal-cortical dysfunction. *Biological Psychiatry*, *62*, 756–764.

Waltz, J. A., Frank, M. J., Wiecki, T. V., & Gold, J. M. (2011). Altered probabilistic learning and response biases in schizophrenia: Behavioral evidence and neurocomputational modeling. *Neuropsychology*, *25*, 86–97.

White, C. (2009). *A model-based analysis of anxiety and biased processing of threatening information*. Retrieved from http://hdl.handle.net/1811/44499

White, C. N., Ratcliff, R., Vasey, M. W., & McKoon, G. (2010). Using diffusion models to understand clinical disorders. *Journal of Mathematical Psychology*, *54*, 39–52.

Wiecki, T. V., & Frank, M. J. (2013). A computational model of inhibitory control in frontal cortex and basal ganglia. *Psychological Review*, *120*, 329–355.

Wiecki, T. V., Sofer, I., & Frank, M. J. (2013). HDDM: Hierarchical Bayesian estimation of the drift-diffusion model in Python. *Frontiers in Neuroinformatics*, *7*, 14. Retrieved from http://journal.frontiersin.org/Journal/10.3389/fninf.2013.00014/full

Windmann, S., & Krüger, T. (1998). Subconscious detection of threat as reflected by an enhanced response bias. *Consciousness and Cognition*, *7*, 603–633.

Zaghloul, K. A., Weidemann, C. T., Lega, B. C., Jaggi, J. L., Baltuch, G. H., & Kahana, M. J. (2012). Neuronal activity in the human subthalamic nucleus encodes decision conflict during action selection. *Journal of Neuroscience*, *32*, 2453–2460.

Zavala, B., Brittain, J.-S., Jenkinson, N., Ashkan, K., Foltynie, T., Limousin, P., . . . Brown, P. (2013). Subthalamic nucleus local field potential activity during the Eriksen flanker task reveals a novel role for theta phase during conflict monitoring. *Journal of Neuroscience*, *33*, 14758–14766.